

Université de Montréal

**Regulation of gene expression in the dinoflagellate *Lingulodinium
polyedrum***

par
Sougata Roy

**Département de Sciences Biologiques
Faculté des Arts and Sciences**

Thèse présentée à la Faculté Faculté des études supérieures
en vue de l'obtention du grade de Ph.D
en Sciences Biologiques

July, 2013

© Sougata Roy, 2013

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée:

Regulation of gene expression in the dinoflagellate *Lingulodinium polyedrum*

Présentée par:

Sougata Roy

a été évaluée par un jury compose des personnes suivantes

.....**Annie Angers**.....

président rapporteur

.....**David Morse**.....

directeur de recherche

.....**Armand Séguin**.....

membre du jury

.....**Frances M. Van Dolah**.....

examineur externe

..... **Annie Angers**

représentant du doyen de la FES

Résumé

Les dinoflagellés sont des eucaryotes unicellulaires que l'on retrouve autant en eau douce qu'en milieu marin. Ils sont particulièrement connus pour causer des fleurs d'algues toxiques nommées 'marée-rouge', ainsi que pour leur symbiose avec les coraux et pour leur importante contribution à la fixation du carbone dans les océans. Au point de vue moléculaire, ils sont aussi connus pour leur caractéristiques nucléaires uniques, car on retrouve généralement une quantité immense d'ADN dans leurs chromosomes et ceux-ci sont empaquetés et condensés sous une forme cristalline liquide au lieu de nucléosomes. Les gènes encodés par le noyau sont souvent présents en multiples copies et arrangés en tandem et aucun élément de régulation transcriptionnelle, y compris la boîte TATA, n'a encore été observé. L'organisation unique de la chromatine des dinoflagellés suggère que différentes stratégies sont nécessaires pour contrôler l'expression des gènes de ces organismes. Dans cette étude, j'ai abordé ce problème en utilisant le dinoflagellé photosynthétique *Lingulodinium polyedrum* comme modèle. *L. polyedrum* est d'un intérêt particulier, car il a plusieurs rythmes circadiens (journalier). À ce jour, toutes les études sur l'expression des gènes lors des changements circadiens ont démontrées une régulation à un niveau traductionnel. Pour mes recherches, j'ai utilisé les approches transcriptomique, protéomique et phosphoprotéomique ainsi que des études biochimiques pour donner un aperçu de la mécanique de la régulation des gènes des dinoflagellés, ceci en mettant l'accent sur l'importance de la phosphorylation du système circadien de *L. polyedrum*.

L'absence des protéines histones et des nucléosomes est une particularité des dinoflagellés. En utilisant la technologie RNA-Seq, j'ai trouvé des séquences complètes encodant des histones et des enzymes modifiant les histones. *L. polyedrum* exprime donc des séquences conservées codantes pour les histones, mais le niveau d'expression protéique est plus faible que les limites de détection par immunodétection de type Western.

Les données de séquençage RNA-Seq ont également été utilisées pour générer un transcriptome, qui est une liste des gènes exprimés par *L. polyedrum*. Une recherche par

homologie de séquences a d'abord été effectuée pour classifier les transcrits en diverses catégories (Gene Ontology; GO). Cette analyse a révélé une faible abondance des facteurs de transcription et une surprenante prédominance, parmi ceux-ci, des séquences à domaine *Cold Shock*. Chez *L. polyedrum*, plusieurs gènes sont répétés en tandem. Un alignement des séquences obtenues par RNA-Seq avec les copies génomiques de gènes organisés en tandem a été réalisé pour examiner la présence de transcrits polycistroniques, une hypothèse formulée pour expliquer le manque d'élément promoteur dans la région intergénique de la séquence de ces gènes. Cette analyse a également démontré une très haute conservation des séquences codantes des gènes organisés en tandem.

Le transcriptome a également été utilisé pour aider à l'identification de protéines après leur séquençage par spectrométrie de masse, et une fraction enrichie en phosphoprotéines a été déterminée comme particulièrement bien adapté aux approches d'analyse à haut débit. La comparaison des phosphoprotéomes provenant de deux périodes différentes de la journée a révélée qu'une grande partie des protéines pour lesquelles l'état de phosphorylation varie avec le temps est reliées aux catégories de liaison à l'ARN et de la traduction. Le transcriptome a aussi été utilisé pour définir le spectre des kinases présentes chez *L. polyedrum*, qui a ensuite été utilisé pour classifier les différents peptides phosphorylés qui sont potentiellement les cibles de ces kinases. Plusieurs peptides identifiés comme étant phosphorylés par la Casein Kinase 2 (CK2), une kinase connue pour être impliquée dans l'horloge circadienne des eucaryotes, proviennent de diverses protéines de liaison à l'ARN.

Pour évaluer la possibilité que quelques-unes des multiples protéines à domaine Cold Shock identifiées dans le transcriptome puissent moduler l'expression des gènes de *L. polyedrum*, tel qu'observé chez plusieurs autres systèmes procaryotiques et eucaryotiques, la réponse des cellules à des températures froides a été examinée. Les températures froides ont permis d'induire rapidement un enkystement, condition dans laquelle ces cellules deviennent métaboliquement inactives afin de résister aux conditions environnementales défavorables. Les changements dans le profil des phosphoprotéines seraient le facteur majeur causant la formation de kystes. Les phosphosites prédits pour être phosphorylés par la CK2

sont la classe la plus fortement réduite dans les kystes, une découverte intéressante, car le rythme de la bioluminescence confirme que l'horloge a été arrêtée dans le kyste.

Mots-clés: dinoflagellé, *Lingulodinium*, expression de gène, RNA-Seq, transcriptome, transcription, traduction, horloge circadienne, histones, kystes, modification post-traductionnelle, kinase, CK2, phosphoprotéomique

Abstract

Dinoflagellates are unicellular eukaryotes found in both marine and freshwater environments. They are best known for causing toxic blooms called ‘red-tides’, for their symbiosis with corals, and for their important contribution to carbon fixation in the ocean. On a more molecular level, they are also known for their unique nuclear characteristics, as they generally have huge amount of DNA found in chromosomes that are permanently condensed and packaged into liquid crystalline forms instead of nucleosomes. Nuclear-encoded genes are often present in multiple copies and arranged in tandem, and no putative promoter elements including the conserved TATA box, have yet been observed. The unique organization of dinoflagellate chromatin suggests different strategies may be required to regulate gene expression in these organisms. In this study, I have started to address this problem using the photosynthetic dinoflagellate *Lingulodinium polyedrum* as a model. *L. polyedrum* is of particular interest because it shows a number of circadian (daily) rhythms. To date, all circadian changes in gene expression studied are regulated at a translational level. I have used transcriptomic, proteomic and phosphoproteomic approaches along with biochemical studies to provide insight into the gene regulatory mechanisms in dinoflagellates, with particular emphasis on the importance of phosphorylation in the *L. polyedrum* circadian system.

The absence of histone proteins and nucleosomes is a hallmark of the dinoflagellates. Using high throughput RNA-seq technology, I found complete set of sequences encoding the core histones as well as sequences encoding histone-modifying enzymes in *L. polyedrum*. Thus *L. polyedrum* expresses conserved histone transcripts, although levels of proteins are still below what can be detected using immunoblotting studies.

Using the *de novo* assembly algorithm the RNA-seq data was used to generate a transcriptome. This transcriptome, a list of genes expressed by *L. polyedrum*, has been extensively characterized. First, homology based sequence searches were used to classify the transcripts in gene ontology (GO) categories, and this analysis revealed a reduced number of

transcription factor types and a surprising predominance of sequences containing a cold shock domain. Alignments of reads from the RNA-seq to genomic copies of *L. polyedrum* tandem repeat sequences was performed to assess the possibility of polycistronic transcripts, a hypothesis proposed to explain the lack of promoter elements in the intergenic region of the tandem repeat gene sequences. This analysis also showed a surprisingly high conservation of tandemly repeated gene sequences.

The transcriptome database was also used to fuel gene identification after protein sequencing by mass spectrometry, and a purified phosphoproteome fraction was found to be particularly amenable to high throughput approaches. A comparison of the phosphoproteome at two different times of day revealed that a major class of proteins whose phosphorylation state varied over time belonged to the RNA binding and translation GO category. The transcriptome was also used to define the spectrum of kinases present in *L. polyedrum*, which in turn was used to classify the different phosphorylated peptides as potential kinase targets. Predicted peptides of casein kinase 2 (CK2), a kinase known to be involved in the circadian clocks of other eukaryotes, were found to include many RNA binding proteins.

To assess the possibility that some of the many cold shock domain proteins identified in the transcriptome might modulate gene expression in *L. polyedrum*, as has been observed in many other eukaryotic and prokaryotic systems, the cellular response to cold temperatures was examined. Cold temperatures were found to induce rapid encystment, a metabolically inactive cell type whose role is to combat unfavourable environmental conditions. Changes in phosphoproteome profile were found to be the major molecular correlates to cyst formation. Predicted CK2 phosphosites are the most highly reduced class of kinase targets, a finding of interest as measurements of the bioluminescence rhythm confirmed that the clock is stopped in cysts.

Keywords: dinoflagellate, *Lingulodinium*, gene expression, RNA-seq, transcriptome, transcription, translation, circadian clock, histones, cysts, posttranslational modification, kinase, CK2, phosphoproteomics

Table des matières

| | |
|---|-------|
| Résumé | iv |
| Abstract..... | vii |
| Table des matières | ix |
| Liste des tableaux | xiv |
| Liste des figures..... | xv |
| List of abbreviations | xvii |
| Dedication..... | xx |
| Remerciements | xxi |
| My Project | xxiii |
| CHAPTER 1-INTRODUCTION | 1 |
| 1.1. The Dinoflagellate | 2 |
| 1.1.1. <i>Lingulodinium polyedrum</i> (previously <i>Gonyaulax polyedra</i>) | 5 |
| 1.1.2. Circadian clocks and <i>L. polyedrum</i> | 6 |
| 1.1.3. Dinoflagellate nuclear genome..... | 9 |
| 1.1.4. Nucleosomes and Chromatin..... | 9 |
| 1.1.5. The Dinoflagellate Cysts | 11 |
| 1.2. Transcription and Maturation of mRNA in Dinoflagellates..... | 15 |
| 1.2.1. Abstract..... | 16 |
| 1.2.3. Transcription and its regulation..... | 18 |
| 1.2.3.1. <i>Cis-acting sequences and RNA polymerase components</i> | 18 |
| 1.2.3.2. <i>Basal/General Transcription factors</i> | 21 |
| 1.2.3.3. <i>DNA binding proteins</i> | 22 |
| 1.2.3.4. <i>Transcriptional regulation</i> | 25 |
| 1.2.4. Splicing and the spliceosome | 31 |
| 1.2.5. RNA transport and mRNA surveillance pathways..... | 34 |
| 1.2.6. Conclusions and perspectives | 36 |
| 1.2.7. Acknowledgments | 50 |

| | |
|---|-----|
| 1.3. Translation in Dinoflagellates | 51 |
| 1.3.1. General Translation | 51 |
| 1.3.2. Translation factors | 52 |
| 1.3.3. Aminoacyl-tRNA synthetases | 54 |
| 1.3.4. Translational regulation..... | 54 |
| 1.3.4.1. <i>By protein factors</i> | 54 |
| 1.3.4.2. <i>By small RNAs</i> | 59 |
| 1.3.5. Posttranslational regulation of gene expression | 61 |
| CHAPTER 2 – PUBLICATION # 1 | 69 |
| A full suite of histone and histone modifying genes are transcribed in the dinoflagellate <i>Lingulodinium</i> | 70 |
| 2.1. Abstract..... | 71 |
| 2.2. Introduction | 72 |
| 2.3. Materials and Methods | 74 |
| 2.3.1. <i>Cell Culture</i> | 74 |
| 2.3.2. <i>Acid Extraction of proteins</i> | 74 |
| 2.3.3. <i>SDS-PAGE and Immunoblotting</i> | 75 |
| 2.3.4. <i>Mass Spectrometric analysis</i> | 76 |
| 2.3.5. <i>Bioinformatic and Phylogenetic Analysis</i> | 77 |
| 2.4. Results | 78 |
| 2.4.1. <i>All core histone and many histone modifying enzyme sequences are present in the Lingulodinium transcriptome</i> | 78 |
| 2.4.2. <i>Phylogenetic grouping identifies at least two major variants of all histone sequences within Lingulodinium</i> | 78 |
| 2.4.3. <i>Histone mRNAs abundance levels are uniform throughout</i> | 79 |
| 2.4.4. <i>Histone protein accumulation is below current detection limits</i> | 79 |
| 2.5. Discussion..... | 81 |
| 2.6. Acknowledgements | 111 |

| | |
|--|-----|
| CHAPTER 3 – PUBLICATION # 2 | 112 |
| Transcripts from dinoflagellate tandem array genes are highly conserved and are not polycistronic | 113 |
| 3.1. Abstract..... | 114 |
| 3.2. Introduction | 115 |
| 3.3. Materials and Methods | 117 |
| 3.5.1. <i>Cell culture</i> | 117 |
| 3.5.2. <i>RNA purification and sequencing</i> | 117 |
| 3.5.3. <i>Sequence assembly and analysis</i> | 117 |
| 3.4. Results | 120 |
| 3.3.1. <i>The de novo assembly is an authentic portrait of the transcriptome</i> | 120 |
| 3.3.2. <i>Tandem gene array sequences are highly conserved in the transcriptome</i> | 122 |
| 3.3.3. <i>Sequences of potential bacterial origin have the same GC-content as the host</i> | 123 |
| 3.3.4. <i>Assessing the potential for polycistronic transcripts</i> | 125 |
| 3.5. Discussion..... | 127 |
| 3.6. Acknowledgements | 156 |
| CHAPTER 4 – PUBLICATION # 3 | 157 |
| Predicted Casein Kinase 2 sites in RNA binding proteins of <i>Lingulodinium</i> show daily variations in phosphorylation state | 158 |
| 4.1. Abstract..... | 159 |
| 4.2. Introduction | 160 |
| 4.3. Materials and methods..... | 163 |
| 4.5.1. <i>Cell Culture</i> | 163 |
| 4.5.2. <i>Phosphoprotein purification and gel electrophoresis</i> | 163 |
| 4.5.3. <i>Mass Spectrometric analysis</i> | 164 |
| 4.5.4. <i>Bioinformatic Analysis</i> | 166 |
| 4.4. Results | 168 |

| | |
|--|-----|
| 4.3.1 Phosphoprotein purification yields more peptides than phosphopeptide enrichment | 168 |
| 4.3.2. The phosphoproteome fraction is enriched in proteins involved in translation and RNA binding | 169 |
| 4.3.3. Phosphopeptide intensity comparisons between ZT2 and ZT14 reveal many RNA binding proteins | 170 |
| 4.3.4. Orthologs of kinases involved in circadian regulation in other eukaryotes may regulate translation in <i>Lingulodinium</i> | 171 |
| 4.5. Discussion..... | 173 |
| 4.6. Acknowledgements | 211 |
| CHAPTER 5 – PUBLICATION # 4 | 212 |
| Cold-induced cysts of the dinoflagellate <i>Lingulodinium</i> have low levels of protein phosphorylation and lack a normal circadian bioluminescence rhythm..... | 213 |
| 5.1. Abstract..... | 214 |
| 5.2. Introduction | 215 |
| 5.3. Materials and Methods | 218 |
| 5.3.1. Cell Culture | 218 |
| 5.3.2. Cyst formation and purification | 218 |
| 5.3.3. RNA Extraction and sequencing..... | 219 |
| 5.3.4. Sequence Analysis..... | 219 |
| 5.3.5. Northern hybridization | 219 |
| 5.3.6. Microscopy | 220 |
| 5.3.7. Protein and phosphoprotein extraction | 221 |
| 5.3.8. 2-D gel electrophoresis | 222 |
| 5.3.9. SDS-PAGE and Western blotting | 223 |
| 5.3.10. Mass spectrometry analysis..... | 223 |
| 5.3.11. Bioinformatic analysis..... | 225 |
| 5.3.12. Bioluminescence assay | 225 |
| 5.4. Results | 226 |
| 5.4.1. Cold temperatures induce temporary cysts in <i>Lingulodinium</i> | 226 |

| | |
|---|-----|
| 5.4.2. <i>Protein phosphorylation is reduced in cysts</i> | 226 |
| 5.4.3. <i>Cysts have an arrested clock and show a decreased level of Casein Kinase 2 phosphosites</i> | 228 |
| 5.4.4. <i>Plastid-encoded RNAs have decreased levels in cysts</i> | 229 |
| 5.5. Discussion..... | 231 |
| 5.6. Acknowledgements | 265 |
| CHAPTER 6 – GENERAL DISCUSSION..... | 266 |
| 6.1. General Discussion | 267 |
| 6.2. Future perspectives | 271 |
| 7. Bibliographie | 273 |

Liste des tableaux

| | |
|--|-----|
| Table 1.2.1. mRNA transport components | 38 |
| Table 1.2.2. mRNA surveillance components | 39 |
| Table 1.3.1. Translational factors in dinoflagellates | 64 |
| Table 1.3.2. Aminoacyl-tRNA synthetases in dinoflagellates..... | 66 |
| Table 1.3.3. Ubiquitin mediated proteolysis..... | 68 |
| Table 2.1. Description of histone sequences and their relative abundance in <i>Lingulodinium</i> .. | 85 |
| Table 2.2. Description of histone modifying enzymes and histone chaperones..... | 86 |
| Table 2.3. Proteins found by LC-MS/MS sequencing of total acid soluble proteins from <i>Lingulodinium</i> and yeast..... | 87 |
| Table 2.4. mRNA abundance of expressed proteins detected by LC-MS/MS in an acid-extracted protein fraction..... | 88 |
| Table 3.ST1. Number of KEGG genes found for a variety of pathways | 151 |
| Table 3.ST2. Number of KEGG pathway sequences found in mammals, plants, apicomplexans, dinoflagellates, ciliates and diatoms for replication, transcription, splicing and translation..... | 152 |
| Table 3.ST3. Nuclear- and plastid-encoded reference sequences from GenBank used for comparison of synonymous and non-synonymous mutations..... | 155 |
| Table 4.1. Comparison of phosphopeptide and phosphoprotein enrichment protocols | 178 |
| Table 4.2. Number of hyperphosphorylated RNA-binding proteins at either ZT2 or ZT14 | 179 |
| Table 4.ST1. <i>Lingulodinium</i> kinases | 194 |
| Table 4.ST2. The identification of proteins containing the 527 phosphopeptides in ZT2 and ZT14 extracts of <i>Lingulodinium</i> | 195 |
| Table 5.ST1. Cyst hyperphosphorylated peptides | 252 |
| Table 5.ST2. Cyst hypophosphorylated peptides | 256 |
| Table 5.ST3. Differential RNA expression in cysts | 263 |

Liste des figures

| | |
|---|-----|
| Figure 1.1.1. The scanning electron microscopy photograph of a single cell of the dinoflagellate <i>Lingulodinium polyedrum</i> taken with a FEI Quanta 200 3D (Dualbeam) microscope..... | 13 |
| Figure 1.2.1. Superphylum Alveolata..... | 40 |
| Figure 1.2.2. <i>Lingulodinium polyedrum</i> nuclear morphology | 42 |
| Figure 1.2.3. RNA polymerase components in <i>Lingulodinium</i> | 44 |
| Figure 1.2.4. TBP phylogenetic classification..... | 46 |
| Figure 1.2.5. General transcription factors in dinoflagellates | 48 |
| Figure 2.1. Two variants of Histone H2A in <i>Lingulodinium</i> | 89 |
| Figure 2.2. The acid soluble protein profiles of <i>Lingulodinium</i> and Yeast differ | 91 |
| Figure 2.3. Histone H3 protein levels in <i>Lingulodinium</i> are below current immunodetection limits | 93 |
| Figure 2.S1. Cladogram of histone H2B | 95 |
| Figure 2.S2. Cladogram of histone H3 | 97 |
| Figure 2.S3. Cladogram of histone H4 | 99 |
| Figure 2.S4. Histone H2B protein is not detected in <i>Lingulodinium</i> | 101 |
| Figure 2.S5. Alignment of H2AX sequences. | 103 |
| Figure 2.S6. Alignment of H2B sequences | 105 |
| Figure 2.S7. Alignment of H3 sequences | 107 |
| Figure 2.S8. Alignment of H4 sequences | 109 |
| Figure 3.1. Global analysis of the <i>Lingulodinium</i> assembly | 131 |
| Figure 3.2. Sequence variation among transcripts..... | 133 |
| Figure 3.3. Bacteria-like sequences in the transcriptomes of different dinoflagellates have GC-contents commensurate with the host..... | 135 |
| Figure 3.4. RNA-Seq does not support a polycistronic transcription mechanism | 137 |
| Figure 3.S1. Size distribution of sequences in the transcriptome and in the mRNA | 139 |
| Figure 3.S2. Degree of sequence identity of <i>Lingulodinium</i> ESTs with the transcriptome | 141 |

| | |
|---|-----|
| Figure 3.S3. Characterization of sequence common to <i>Lingulodinium</i> , <i>Alexandrium</i> and <i>Karenia</i> | 143 |
| Figure 3.S4. Frequency spectrum of sequence variation in PCP and Luc TAG transcripts.... | 145 |
| Figure 3.S5. Characteristics of the bacterial-like sequence in the transcriptome | 147 |
| Figure 3.S6. Detection of reads aligning to the Luciferase intergenic spacer | 149 |
| | |
| Figure 4.1. Efficient enrichment of <i>Lingulodinium</i> phosphoproteins by affinity chromatography | 180 |
| Figure 4.2. RNA rather than DNA related processes are preferred in the total enriched protein pool..... | 182 |
| Figure 4.3. Phosphopeptide Intensity at ZT2 is much pronounced than at ZT14 | 184 |
| Figure 4.4. Many RBP are differentially phosphorylated at ZT2 and ZT14 | 186 |
| Figure 4.5. Daily variation of kinase activity and their efficiency in <i>Lingulodinium</i> | 188 |
| Figure 4.6. Many RNA binding proteins are among the predicted CK2 targets | 190 |
| Figure 4.S1. Comparison of kinases | 192 |
| | |
| Figure 5.1. Cyst morphology differs from that of motile cells..... | 236 |
| Figure 5.2. Phosphoprotein profiles of cyst and motile cell extracts differ..... | 238 |
| Figure 5.3. Cyst phosphopeptides are generally hypophosphorylated and fall into categories regulating the amount and activity of proteins | 240 |
| Figure 5.4. Casein Kinase 2 phosphosites are the most hypophosphorylated class | 242 |
| Figure 5.5. RNA-Seq of cyst extracts reveals most RNAs with altered levels have decreased abundance | 244 |
| Figure 5.S1. Western blot analysis of three nuclear encoded proteins show no significant decrease in cyst extracts. | 246 |
| Figure 5.S2. Comparison of phosphoprotein enrichment fraction from cyst and motile cells | 248 |
| Figure 5.S3. Northern blot analyses of two plastid-encoded and two nuclear-encoded RNAs confirms a decrease in plastid RNAs..... | 250 |

List of abbreviations

2D-PAGE: 2-dimensional polyacrylamide gel electrophoresis

AMPK: 5' adenosine monophosphate-activated protein kinase

ATP: adenosine 5'-triphosphate

bp: base pair

BSA: bovine albumin

cDNA: complementary DNA

CDPK: calcium-dependent protein kinase

CHAPS: 3-[(3-Cholamidopropyl) dimethylammonio]-1-propanesulfonate hydrate

CK: casein Kinase

CSD: cold shock domain

CT: circadian time

DBP: DNA binding protein

Dinap1: dinoflagellate nuclear associated protein 1

Dip: Dinap interacting protein

DNA: deoxyribonucleic acid

DTT: dithiothreitol

EDTA: ethylenediaminetetraacetic acid

GAPDH: glyceraldehyde-3-phosphate dehydrogenase

GDP: guanosine 5'-diphosphate

GSK: glycogen synthase kinase

GTP: guanosine 5'-triphosphate

HEPES: N-(2-Hydroxyethyl) piperazine-N'-(2-ethanesulfonic acid) hemisodium salt

| | |
|---------------|---|
| IEF: | isoelectric focussing |
| Kb: | kilobase |
| kDa: | kilodalton |
| KH: | K-homology |
| LB: | luria-Bertoni |
| LBP: | luciferin binding protein |
| LCF: | luciferase |
| LD: | Light/Dark |
| LDS: | lithium dodecyl sulfate |
| mg: | miligram |
| miRNA: | microRNA |
| mL: | mililiter |
| mM: | millimolar |
| MMG: | monomethylguanosine |
| MOPS: | 3-(N-Morpholino propanesulfonic acid |
| mRNA: | messenger RNA |
| MS: | mass spectrometry |
| NADP: | nicotinamide adenine dinucleotide phosphate |
| nm: | nanomolar |
| OD: | optical density |
| ORF: | open reading frame |
| PCP: | peridinin Chlorophyll-a binding protein |
| PCR: | polymerase chain reaction |
| PEG: | polyethylene glycol |
| PFAM: | protein families |

| | |
|--------------|---|
| PMSF: | phenylmethanesulfonyl fluoride |
| PPR: | pentatricopeptide |
| PSI: | photosystem I |
| PSII: | photosystem II |
| PTM: | posttranslational modification |
| RBP: | RNA binding proteins |
| RNA: | ribonucleic acid |
| RNAi: | RNA interference |
| RRM: | RNA Recognition motif |
| rRNA: | ribosomal RNA |
| SDS: | sodium dodecyl sulfate |
| SEM: | scanning electron microscopy |
| SL: | splice leader |
| TBP: | TATA-box binding-protein |
| TLF: | TATA-box like factor |
| TMG: | trimethylguanosine |
| Tris: | 2-amino-2-hydroxyméthyl-1,3-propanediol |
| tRNA: | transfer RNA |
| TRP: | TATA-box related proteins |
| TTFL: | transcription translation feedback loop |
| UTR: | Untranslated region |
| UV: | ultra violet |
| ZT: | Zeitgeber time |

Dedication

*I want to dedicate this thesis to my (Mother)
Ma- my epitome, with her loving and caring
attitude has been always by my side and my
(Father) Baba, my philosopher and guide has
been a source of inspiration to me.*

Remerciements

I would like to convey my deep respect and heartfelt gratitude to Prof. David Morse, my research director, whose constructive criticism and methodical problem solving approach helped me to develop my scientific outlook. I admire his ability as a teacher to captivate the audience with his enthusiastic, energetic and informative explanation. I am thankful to Prof. Morse for the liberty and encouragement he has offered me to pursue several projects. It has been an extraordinary experience to work with Prof. Morse and I am grateful to him for giving me this opportunity. He is one of the most fun-loving and knowledgeable people I have known.

I am grateful to my PhD committee, Prof. Daniel P. Matton and Dr. Jean Rivoal for their helpful suggestions and insightful discussion. I would like to take this opportunity to thank Prof. Mario Cappadocia for his support and guidance during the sabbatical period of Prof Morse. I would like to extend my gratitude to all the faculty members, students and administrative staff of IRBV for their help.

I consider myself lucky to be able to work with very intelligent, passionate and helpful co-researchers Steve and Mathieu. Working with them has been an excellent and pleasant experience, which cannot be explained by mere words. I would also like to thank Philippe for his help during my early days of PhD. Thank you all for making my stay in the Morse Lab memorable. Thanks to Jonathan for the elaborate and interesting discussions we had on different scientific as well as extracurricular topics.

For the complete duration of my PhD, I have always felt the warmth of a family although I was far away from my home. Thanks to Leo and Lise and their sons Etienne and Julien, with whom I have been staying for the last five years. Thank you all for including me as one of the members of your lovely family.

Some debts we can never think of repaying and indebtedness to family comes among the foremost ones. I have no words of appreciation to describe the contribution of my family. Without the continuous and unconditional love, encouragement and guidance of my mother

and father, I would not have come so far in my career. I would like to thank my elder brother, who has always remained by my side inspiring and motivating me. I appreciate the calm presence of my sister-in-law and my special thanks to the newcomer in my family, my niece, Srita.

Coming to Canada and completing my studies would not have been possible without the financial help through grants and scholarships from the IRBV, FESP, the Department of Biological Sciences and the NSERC grant funding our laboratory. Thanks to all the organizations for their generous support.

My Project

My curiosity about *Lingulodinium* was impelled by the observation that gene expression can occur rhythmically, under the control of an internal daily (circadian) clock, even though this species has the permanently condensed chromatin characteristic of the dinoflagellates. My particular interest was to try to understand expression of genes present in multiple copies and arranged in tandem arrays, and I began these studies by analysis of a transcriptome prepared from high throughput sequencing (RNA-Seq) data. During these studies I noticed that the transcriptome contained all core histone genes, as well as many histone modifying enzyme genes. The fact that these sequences were conserved and expressed thus argues against the dogma that histones have been lost in dinoflagellates. However, I was unable to show the presence of histones immunologically, indicating that protein levels are still below the limits of what can be detected experimentally. I also became interested in the study of gene expression at the post-transcriptional level. In particular, I was interested in using the transcriptome database to allow a mass spectrometry protein sequencing approach. I found that an enriched phosphoprotein fraction could be directly analyzed. Phosphoproteomics is fast emerging as an important field to study the role of gene expression at a posttranslational level. The *Lingulodinium* phosphoproteome shows important differences between day and night suggesting that protein phosphorylation may be a mechanism whereby the *Lingulodinium* circadian system influences cell behavior. Lastly, the phosphoproteome is markedly different in cold-induced cysts, which together with the observation that the circadian clock has stopped in cysts, may help to identify components of the clock machinery.

CHAPTER 1-INTRODUCTION

1.1. The Dinoflagellate

Dinoflagellates are unicellular eukaryotic protists found widely in marine as well as freshwater environments. The fossil record reveals traces of dinoflagellates 245 million years ago, confirming their presence during the Mesozoic and Cenozoic era. However, molecular phylogenetic studies coupled with anatomical comparison suggests that origin of the modern dinoflagellates may actually be found in the Precambrian, more than 570 million years ago [1]. Sequence analyses and the construction of phylogenetic trees place dinoflagellates within the kingdom Alveolata, which also contains the parasitic group of apicomplexans and ciliates [2-4], which are unified by the presence of large flattened cortical vesicles termed alveoli. Dinoflagellates themselves are further sub-divided into oxyrrhinales, syndinales and the core dinoflagellates groups [5]. Dinoflagellates are generally microscopic, ranging from 15-50 μM in size, with the largest reported dinoflagellate, *Noctiluca*, up to 2 mm in diameter [6-8]. Some dinoflagellates, termed armored, contain a series of cellulosic thecal plates inside the alveoli [9], although unarmored species lacking the thecal plates also exists [10]. The dinoflagellates demonstrate extensive diversity with respect to morphology and food habits with photosynthetic [11], heterotrophic [12], symbiotic [13], parasitic [14] or mixotrophic [15] behavior all found in the group. A large number of the dinoflagellates are photosynthetic, and most contain distinctive plastids surrounded by three membranes, thought to result from a secondary endosymbiosis, as well as containing the unique xanthophyll peridinin, a carotenoid responsible for the characteristic red color in these species [16, 17]. The peridinin-containing dinoflagellates have the smallest chloroplast genome of any functioning plastid because of frequent gene transfer events from the chloroplast to the nucleus [18]. However, other types of plastid are occasionally found in some dinoflagellate species such as the fucoxanthin containing dinoflagellates *Karenia* and *Karlodinium*. These organisms are thought to have acquired their chloroplasts during a tertiary endosymbiosis, where the peridinin-containing plastids have been replaced by a fucoxanthin-containing plastid [19, 20]. Dinoflagellates also have a characteristic helical swimming behavior due to its two flagella, one running like a belt in a transverse groove around the cell, which acts to spin the cell around its axis, the other lying in a longitudinal groove toward the base of the cell which pushes the cell forward [21, 22]. Dinoflagellates

can decide the direction of their movement by sensing chemicals, light as well as gravity [21].

Another interesting feature of dinoflagellates is its capability to form symbioses with different marine protists and invertebrates such as foraminifera, radiolarians, flatworms, anemones, jellyfish, and mollusks [23-25]. In particular, the association of *Symbiodinium* with the reef-forming corals has been of great interest because of the immense importance reefs play in marine ecology [25, 26]. In this mutual relationship, which benefits both the organisms, the photosynthetic dinoflagellates fix atmospheric CO₂ and transfer a considerable part of it as food for the corals [27] in exchange of shelter.

Toxic dinoflagellates can have a huge negative impact on different forms of marine life [27-30] and on human [31] through the formation of harmful algal blooms (HABs). HAB outbreaks are becoming more frequent in the coastal areas around the world, and at least in part, are due to the increase of nitrogen and phosphate (fertilizer) runoff in the coastal waters [32]. HAB-causing dinoflagellates can release different types of toxins, the most studied of which are neurotoxins, resulting in various syndromes and can even be fatal to humans. The most common toxin, saxitoxin, is a sodium channel blocker that causes paralytic shellfish poisoning (PSP), and is produced mainly by *Alexandrium* species and to some extent by *Gymnodinium catenatum* [33, 34] and *Pyrodinium bahamense* var *compressum* [35]. Brevetoxins, released by *Karenia brevis* [36], also cause the neurotoxic shellfish poisoning (NSP) and marine mammals have been found to be highly susceptible to ingestion or even inhalation of this toxin. *Gambierdiscus toxicus* produces the polyether-based ciguatoxins [37] which cause ciguatera fish poisoning (CFP), while diarrhetic shellfish poisoning (DSP) and azaspiracid shellfish poisoning (AZP) are caused by toxins from *Dinophysis* or *Prorocentrum* species and *Protoperidinium crassipes* [38], respectively.

Horizontal gene transfer (HGT), also known as lateral gene transfer, is the exchange of genetic material between two organisms. HGT can either occur between prokaryotes and eukaryotes or within prokaryotes or eukaryotes themselves [39]. HGT has been reported in many eukaryotes and the genome-wide investigations are now being used to identify the

novel genes introduced into genomes from different sources [40-42]. The use of phylogenomics approaches in particular [43, 44] are extremely useful in revealing the extent of gene transfer from bacteria to protists [45], with up to 7.5% of the genes in diatoms derived from bacterial sources by HGT [45]. In dinoflagellates, many genes such as the histone-like proteins (HLPs), the form II Rubisco, *aroB* and O-methyltransferase have been acquired from bacteria through HGT [46]. Proteorhodopsins of marine proteobacteria have been horizontally transferred to other prokaryotic classes, and recently have also been found in two different dinoflagellates, a result of independent HGT events [47]. The full extent of HGT events has not yet been unveiled for dinoflagellates.

Bioluminescence is a fascinating process found in at least 30 different systems, belonging to phylogenetic groups as diverse as bacteria, fungi, dinoflagellates or insects. The general mechanism of light production involves the oxidation of a substrate (luciferin) by molecular oxygen and is catalyzed by a Luciferase enzyme specific to each system. The marine environment seems to be preferred habitat for bioluminescent species [48]. The dinoflagellates constitute one of the most common sources of bioluminescence in marine waters [49] and at least 18 genera of dinoflagellates are known to produce bioluminescence [50]. However it is curious that among many marine bioluminescent organisms only dinoflagellates can perform photosynthesis [51]. Particularly, the dinoflagellate *Lingulodinium polyedrum* (previously *Gonyaulax polyedra*) has been used as the model to understand the biochemistry, cell biology, and molecular biology of bioluminescence. In the *Lingulodinium* system the luciferin is a linear tetrapyrrole structurally related to chlorophyll, which is highly susceptible to nonluminescent autooxidation. At cytoplasmic pH (~7.5), luciferin in the cell is sequestered by a Luciferin Binding Protein (LBP), preventing its reaction with Luciferase (LCF). Mechanical, chemical or temperature stimulation of the organism is followed by emission of light as brief (~100 msec), bright (~10⁹ photons/cell) flashes from small (~0.4 microns) discrete spherical organelles called scintillons. These organelles are essentially protein spheres, which protrude into the vacuole and are almost completely surrounded by the vacuolar membrane. Stimulation of the cell activates a voltage-gated proton channel in the vacuolar membrane, thus decreasing the cytoplasmic pH in the region of the scintillons (to ~6.5). At this acidic pH, LBP is inactivated and releases

luciferin, while LCF becomes activated at low pH and initiates the bioluminescent reaction. The number of scintillons at night is roughly tenfold higher than at day and corresponds to what has been termed the bioluminescent capacity of the cells. Interestingly, light is also emitted as glow, a constant low intensity ($\sim 10^4$ photons/sec/cell) emission rising to maximum peak at a very specific time each day. The mechanism of glow, not visible to the naked eye, is still unknown. Furthermore, the exact purpose of dinoflagellate bioluminescence is still being debated, although probable functions could be to scare predators away by producing brilliant light [52-54] or to draw the attention of secondary predators to reduce the primary predator population, which is popularly known as the “burglar alarm” effect [55-57].

1.1.1. *Lingulodinium polyedrum* (previously *Gonyaulax polyedra*)

L. polyedrum is a marine, thecate, non-toxic, bioluminescent and photosynthetic dinoflagellate, approximately 35 X 45 μm in size (Figure 1.1.1) [58]. It is one of the species involved in the formation of ‘red tides’, the well-known algal blooms in nutrient rich waters [59]. Though *L. polyedrum* is generally considered as a non-toxic species, and while it has been maintained in culture without incident for over 50 years, some studies were able to detect very low levels of yessotoxins [60]. It contains a triple membrane bound, peridinin-containing chloroplast. *L. polyedrum* has a C-shaped nucleus with a characteristic nucleolus and about 200 pg of DNA distributed over roughly 200 chromosomes [61]. It also contains an electron-dense, spherical PAS body bounded by a single membrane, a structure considered equivalent to the digestive granules found in hydra or food vacuoles of *Ceratium hirundinella* [58]. *L. polyedrum* mitochondria are surrounded by two membranes and have tubular cristae. *L. polyedrum* is a useful model system for understanding the biochemical bases of biological rhythms, as many physiological activities in this alga have daily rhythms that are independent of external cues and are thus regulated by a daily (circadian) clock [62].

1.1.2. Circadian clocks and *L. polyedrum*

For better adaptation to the surrounding environment, living organisms frequently prefer to perform some biological tasks at particular times of day [63]. They are able to do so with the help of an endogenous clock, termed as the circadian clocks, which not only allows the organism to synchronize their internal biochemistry to the daily cues but also to predict the changes to come [64, 65]. While the clock can function in the absence of daily cues, the clock is typically aligned to, or synchronized with the environment by changes in light and/or temperature cues [66, 67]. This mechanism thus requires an efficient signaling pathway to link the external stimuli to the central oscillator, which then sends timing signals to regulate cellular physiology. The circadian system can thus be thought of as being composed of three components, the inputs, the clock itself, and the outputs, known as overt rhythms and often involving changes in gene expression [64]. Extensive research in mammals, plants, insects, fungi and cyanobacteria has characterized the molecular components of the central oscillator and elements responsible for sensing the input and propagating the rhythmic output. Except cyanobacteria, whose core oscillator can function using only phosphorylation/dephosphorylation of clock proteins [68], all other tested organisms have a mechanism involving transcriptional/ translational feedback loops (TTFL) [69, 70]. In its most simple terms, the feedback loop is comprised of transcriptional activators that activate the expression of repressor proteins, which then inhibit the activators thereby closing the loop [70]. This basic mechanism of circadian regulation is conserved across diverse phyla, although the clock proteins themselves are not, which is indicative of distinct evolutionary origins for the clocks in different organisms. Concerning the clock outputs, transcriptional regulation has been the focus of interest for several years, but after the discovery of the cyanobacterial clock mechanism, efforts were directed to assess the importance of posttranscriptional gene regulation in clock functioning. In this context, the dinoflagellate *Lingulodinium polyedrum* has again proved to be an interesting model, as in most observed cases it is found that *Lingulodinium* circadian clock prefers translational rather than the transcriptional regulation of gene expression [71].

The complexity of single cell circadian systems has been well documented in *L. polyedrum*. Many important characteristics of the circadian systems such as the mechanism of temperature compensation [72], basic features of phase shifting by light and the first phase response curve [73], the action spectrum for light phase shifting [74], phase shifting by drugs as well as drug effects on period [75] have been documented from studies performed on *Lingulodinium*. However, as yet, the core clock proteins are not known in any dinoflagellate. One of the major effects of the clock in *Lingulodinium* is to regulate the synthesis of different proteins. A combination of 2-D PAGE and tandem mass spectrometry revealed 28 out of 900 proteins in *L. polyedrum* were expressed differentially at three different phases over the circadian cycle [76]. Many of these proteins have been investigated in an attempt to link them to one of the several rhythmic processes, which include bioluminescence, photosynthesis, cell aggregation and timing of cell division. Indeed, *Lingulodinium* has been extensively researched for the last 60 years to try to understand the biochemical mechanism of circadian clock regulation of these physiological rhythms. Among the different rhythmic outputs, bioluminescence has received particular attention because of the ease of its measurement and as light emission appears to require only two proteins (LBP and luciferase) and the substrate luciferin, and thus constitutes a relatively simple system to study the molecular biology and biochemistry of the links between circadian clocks and the observed rhythms they control. Both LBP and LCF demonstrate robust daily rhythms in protein expression, and both are more abundant during night corresponding to the time of maximum bioluminescence [77-79]. Curiously, the dinoflagellate *Pyrocystis lunula*, which also demonstrates nightly bioluminescence, does not contain LBP and its LCF protein levels are constant throughout the circadian cycle. In this case the rhythmic bioluminescence correlates instead with translocation and compartmentalization of LCF at different times [80], a totally different mechanism from that observed in *Lingulodinium*. This suggests that various different posttranslational strategies may be employed by dinoflagellate circadian clocks to regulate cell physiology.

The role of posttranslational modifications of proteins, and phosphorylation in particular, has been extensively studied in circadian biology. For example, the rhythmic phosphorylation of clock proteins in the cyanobacteria *Synechococcus*, a process that

continues even *in vitro*, is sufficient for the circadian oscillator to function independent of transcription or translation [81]. In *Arabidopsis*, transcript abundance for several kinases and phosphatases were found to be rhythmic, and in addition, some of the core clock proteins (morning-expressed MYB-like transcription factors CIRCADIAN CLOCK ASSOCIATED1 (CCA1) and LATE ELONGATED HYPOCOTYL (LHY) are known to be phosphorylated by casein kinase 2 (CK2) [82, 83]. In *Neurospora*, phosphorylation of the clock protein Frequency (FRQ) determines its rate of degradation, which in turn determines the period length of the circadian oscillator [84]. Similarly, posttranslational regulation of clock components is important for other animal and mammalian circadian clocks [85, 86]. Interestingly, both CK2 and casein kinase 1 delta/epsilon (CK1 δ/ϵ) appear to have a conserved role in the function of eukaryotic circadian clocks [87, 88]. Mutational analysis assays using specific kinase inhibitors and disruption of CK1 δ/ϵ and CK2 genes revealed an arrhythmic behaviour, similar to that observed when core clock proteins were mutated [82, 83, 87, 89-92]. However, it should be noted that other kinases, such as adenosine monophosphate-activated protein kinase (AMPK) or Glycogen synthase kinase -3 β (GSK-3 β), have also been implicated in eukaryotic clocks.

In *Lingulodinium*, the kinase inhibitors staurosporine and 6-dimethylaminopurine induced concentration-dependent lengthening of the free-running period of the bioluminescence rhythm and high concentrations induced complete stoppage of the clock [93, 94]. As the inhibitors used in these experiments were broad range and affect many kinases, it is possible that there are many kinases involved in the *Lingulodinium* circadian clock. The phosphatase inhibitors okadaic acid, calyculin A, and cantharidin, all inhibitors specific for protein phosphatases 1 and 2A also affect the clock but the effects are less pronounced. High concentrations of okadaic acid produce a significant lengthening of the bioluminescent glow rhythm period, whereas only phase delays and no persistent effect on period were observed with cantharidin and calyculin treatments [95, 96]. These studies underscore the importance of phosphorylation/dephosphorylation events in the *Lingulodinium* circadian system, but as yet, the kinase repertoire in *L. polyedrum* has not been characterized.

1.1.3. Dinoflagellate nuclear genome

The dinoflagellates as a class are proof that genome size is not an indicator of an organism's complexity, as different species show a remarkable variability in their genome content. Among the known dinoflagellates, the 3 pg haploid genome of *Symbiodinium* [97] is the closest in size to the 3.2 pg present in a haploid human cell, yet the 250 pg of DNA in *Prorocentrum micans* represents almost a hundred-fold greater genome size [98]. *L. polyedrum* cells contain 200 pg of DNA, about 60 times more than a haploid human cell. Genome size has been linked to the gene content using the positive correlation between the number of genes in an organism and its genome size [99, 100], and the regression model predicted a gene content for the dinoflagellates of around 37000 – 87000 [101]. However, genes in *Lingulodinium* are generally present in several copies (copy numbers from 30 - 5000) suggesting the number of unique genes is bound to be less than these estimates. A *de novo* assembly of 454 sequencing reads for two strains of *Symbiodinium* identified about 55,000 contigs in each species [102], although these assemblies are biased towards shorter fragments and many are likely to be derived from the same gene. Independent from the actual number of genes, it will be very interesting to find out why such simple eukaryotes accumulate such huge amounts of DNA and how they manage to conserve and express the relevant sequences. This is especially relevant given the unusual structure of the dinoflagellate chromatin, sufficiently distinct from other eukaryotes or prokaryotes to have at one time been termed a dinokaryon [103]. The dinoflagellate transcriptomes are generally biased towards high GC content, with an average varying from 50% in *Karlodinium micrum* to about 68% in *Lingulodinium polyedrum*.

1.1.4. Nucleosomes and Chromatin

The DNA in the nucleus must be tightly packed in order to fit within the available space, but more importantly, must also be accessible for replication and transcription. Most eukaryotes fold and compact DNA into chromatin using nucleic acid-protein complexes, the fundamental unit of chromatin is termed as nucleosomes. The nucleosome is a small cylindrical molecule containing 147 base pairs of DNA folded around four pairs of the

highly basic histone proteins H2A, H2B, H3 and H4. These nucleosomes can be further organized with respect to one another using a linker histone H1 protein [104]. Structurally and functionally, the four core histone proteins all have two distinct domains, the first a 20-35 N-terminal extension termed the histone tail and the second a conserved 80-90 C-terminal region termed the histone fold [105]. The fold regions of the H2, H3 and H4 have poor sequence conservation but demonstrate a high degree of structural similarity [106]. This fold domain is responsible for interaction with other histones [105] and DNA [107], and it is this region that determines the structural basis of the nucleosome. Given the balance needed between compactness and accessibility of the DNA, it might be expected that nucleosomes can affect chromatin structure. Indeed, the regulation of transcription through modulation of chromatin structure, popularly known as the epigenetic regulation of gene expression, is now an important field of research. Histone modifications form a part of the epigenetic toolkit. The residues present in the unstructured N terminal histone tails can undergo several types of post-translational modification (e.g acetylation, phosphorylation, methylation, biotinylation, SUMOylation, ADP ribosylation and ubiquitination) and can either activate or inhibit transcription depending on the modification [108]. In addition, some variants of the canonical core histones also exist in eukaryotes and these can have important implications in regulating expression of specific genes [109]. One such example is H2A.Z, a variant of H2A, which when present in the nucleosomes remodels it to affect gene expression [110].

Remarkably, dinoflagellates are the only eukaryotes that do not appear to use nucleosomes to organize their chromatin. Nucleosomes have never been observed microscopically in the chromatin, and none of the conserved histone proteins have ever been detected in dinoflagellate extracts. Instead of histones, the basic protein fraction isolated from dinoflagellate nuclei contains a class of histone-like proteins (HLP) structurally similar to bacterial DNA-binding proteins of the HU type [111, 112]. HLPs do appear able to condense DNA [113], which is in agreement with a predicted role in modulating chromosome structure [114]. Furthermore, some HLPs were found to be post-translationally modified, indicating a possible role in regulating gene transcription [115], but this has not been confirmed experimentally. However, the amount of these basic proteins in dinoflagellate nuclei is extremely low, with a protein/DNA ratio roughly one tenth of what is

found in other eukaryotes. This low level of nucleoproteins is thought to favor the liquid crystalline form of DNA, a type of DNA packing also found in animal sperm cell nuclei, which are also devoid of histones and instead contain a small (~7 kDa), highly arginine rich nucleoproteins known as protamines. Furthermore, imaging of the chromatin with the electron microscope supports a liquid crystalline organization for dinoflagellate chromosomes [116]. Surprisingly, however, recent transcriptomic sequencing studies have shown that conserved histone transcripts were found to be expressed in dinoflagellates [117]. It is difficult to reconcile the presence of these conserved sequences with the lack of detectable protein, although it is possible that histones at low levels will eventually be found to play a role in some aspects of chromatin organization.

In addition to the histone modifications, DNA methylation at the C5 position of cytosine pyridine ring is another epigenetic mechanism that can alter gene expression. In eukaryotes, the methylation catalyzed by DNA methylases generally occurs at CpG dinucleotides on double stranded DNA and is usually associated with gene inactivation [118]. These so-called CpG islands are found frequently in the promoter regions of genes and it has been proposed that DNA binding factors which recognize this structural change can modify gene expression [119]. Dinoflagellate DNA typically contains a considerable amount of modified bases, with roughly 3% of the cytosine replaced by 5-methyl cytosine. This does not appear to be located in CpG islands but instead appears to be distributed randomly throughout the genome [120]. However, the most important modification is replacement of up to 70% of the thymine by 5-hydroxymethyluracil [121]. Unfortunately, in the absence of genome sequence information it is difficult to assess the functional role of such modifications (if any at all).

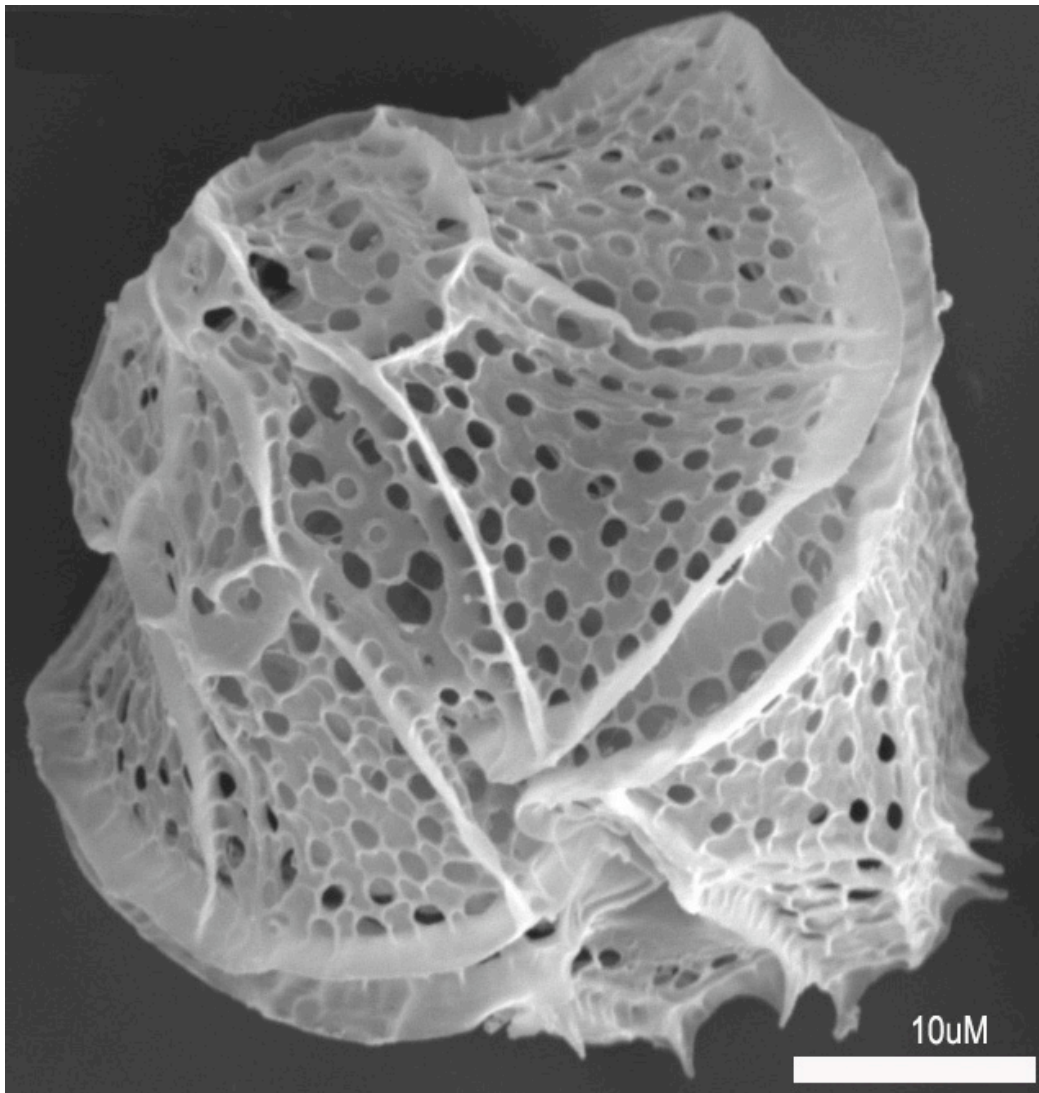
1.1.5. The Dinoflagellate Cysts

The dinoflagellates have left an abundant trail in the fossil record in particular due to their ability to form cysts in order to resist unfavorable conditions. Roughly 10% of the known dinoflagellate species are able to produce resting cysts [122], a dormant cell that can germinate back to form a motile cell when conditions become favourable [123]. .

Ecologically, resting cysts are important as they can act as a “seed-bank” for the formation of HABs [124]. Indeed, many toxic HAB species undergo formation of resting cysts as a part of their life cycle [125]. The permanent cysts can be transported via sediments and remain viable for hundreds of years.

Dinoflagellates can also form temporary cysts, the structures of these are quite different from permanent cysts. In contrast to resting cysts, encystment and germination of temporary cysts can be very rapid, which presumably helps prevent huge population losses in response to a sudden unfavorable condition. Several factors, including temperature [126], chemicals [127], population pressure [128], nutrient deficiency, bacteria [129], pH and salinity [130] are well known inducers of temporary cysts. Interestingly, unlike other eukaryotes and bacteria, some dinoflagellates survive low temperature periods by formation of temporary cysts. This quick response requires an efficient signaling cascade, and the calcium concentration via the phospholipase C pathway has been shown to be vital for dinoflagellate temporary cyst formation in response to melatonin [127]. However, the major molecular events associated with the structural rearrangement and the quiescence physiology is still unknown.

Figure 1.1.1. A scanning electron microscopy photograph of a single cell of the dinoflagellate *Lingulodinium polyedrum* taken with a FEI Quanta 200 3D (Dualbeam) microscope.



1.2. Transcription and Maturation of mRNA in Dinoflagellates

This part of the thesis is under review for publication as a review chapter in the journal *Microorganisms* (ISSN 2076-2607) published by MDPI. This chapter is submitted for peer-review with the title: *Transcription and Maturation of mRNA in Dinoflagellates*.

Sougata Roy¹ and David Morse^{1,*}

¹ Institut de Recherche en Biologie Végétale, Département de Sciences Biologiques,
Université de Montréal, 4101 Sherbrooke est, Montréal, Québec, Canada H1X 2B2

This manuscript was reviewed and corrected by my supervisor (David Morse) after I completed the initial drafting.

1.2.1. Abstract

Dinoflagellates are of great importance to the marine ecosystem, yet scant details of how gene expression is regulated at the transcriptional level are available. Transcription is of interest in the context of the chromatin structure in the dinoflagellates, which shows many differences from more typical eukaryotic cells. Here we canvas recent transcriptome profiles to identify the molecular building blocks available for the construction of the transcriptional machinery and contrast these with those used by other systems. Dinoflagellates display a clear paucity of specific transcription factors, although surprisingly, the rest of the basic transcriptional machinery is not markedly different from what is found in the close relatives to the dinoflagellates.

Keywords: transcription factor; gene expression; regulation

1.2.2. Introduction

Dinoflagellates are an important group of unicellular eukaryotes found in both marine and fresh water environments. These marine species are of particular importance on a global scale, as along with the diatoms, they contribute roughly half of the carbon fixed in the oceans, and thus roughly a quarter of the global totals [131]. They also play a role in maintaining the biodiversity surrounding coral reefs, since the coral polyps themselves rely on photosynthetic products supplied by the symbiotic dinoflagellates they harbor for growth in nutrient poor waters [132]. Furthermore, many marine dinoflagellates synthesize potent toxins that accumulate to high concentrations in the algal blooms commonly called “red tides” [133]. Lastly, the nightly bioluminescence of many dinoflagellates, popularly known as the “phosphorescence of the sea”, has inspired not only art and literature but also intensive scientific dissection of the bioluminescence phenomenon [134]. Interestingly, in *Lingulodinium polyedrum* this nightly bioluminescence [135], as well as photosynthesis [136], cell division [137], and diurnal vertical migration [138], are all regulated by an endogenous circadian (daily) clock. *L. polyedrum* has been studied for over 60 years as a model system for addressing the biochemical links between the internal clock and the observed rhythms [71].

Phylogenetically, dinoflagellates are grouped in the superclass Alveolata, which contains apicomplexans as their closest relatives as well as ciliates [4]. This group has several features unique to these organisms (Figure 1.2.1.). However, dinoflagellates have many unique characteristics compared to their relatives. For example, dinoflagellates typically possess a large quantity of nuclear DNA containing many genes organized in tandem gene arrays, with DNA found in a liquid crystal structure lacking observable nucleosomes [61]. It is unfortunate that dinoflagellates have so far proven refractory to mutational or gene transformational studies, thus hindering extensive molecular studies, as these unusual nuclear features suggest that the mechanisms for gene expression and its regulation may also be unusual.

The mechanisms used to control the expression of different genes have been extensively researched in both prokaryotes and eukaryotes. Critical events in eukaryotes include changes in chromatin organization, transcription of DNA into pre-mRNA, splicing of pre-RNA into mature mRNA, mRNA transport, mRNA degradation, mRNA editing and covalent modifications of the mRNA, translation of mRNA into protein, and, lastly, post-translational modification of the protein. All these, either individually or collectively are responsible for regulating gene expression within a cell. In this review we will focus primarily on transcription and its regulation as they relate to the control of gene expression in the dinoflagellates.

1.2.3. Transcription and its regulation

1.2.3.1. Cis-acting sequences and RNA polymerase components

Dinoflagellate chromosomes are permanently condensed at all stages of the cell cycle (Figure 1.2.2.) and assume a liquid crystalline structure [139, 140] with bivalent cations acting as the stabilization matrix [141]. This unusual chromatin structure thus raises the important questions about the accessibility of genes within the structure to the transcriptional machinery. The dinoflagellate *Prorocentrum micans* was inspected using high resolution electron microscope autoradiography for ³H-adenine incorporation, and this revealed that RNA transcription was prevalent only on extrachromosomal DNA filaments and not on DNA within the main body of the chromosome [142]. It was proposed that this transcriptionally inactive DNA might instead play a role in stabilizing chromosome organization, perhaps by an association with a protein matrix [142].

Given access to the genetic material, transcription initiation in dinoflagellates is likely to require an elaborate set of *trans*-acting factors and a series of conserved *cis*-acting sequences, as is the case in other eukaryotes. The complex of trans-acting factors binding the regulatory sequences in the DNA includes, in addition to the RNA polymerases, both general and gene-specific transcription factors, activators and mediators [143]. The *cis*-acting

sequences in eukaryotes can include regulatory elements far from the transcription start site, termed enhancers, although the region just upstream of the start site, termed a promoter, consisting of a core region and other regulatory domains [144, 145] is considered as the primary site of initiation. There are two major classes of promoters that regulate the expression of protein coding genes, and these contain either a TATA-box (consensus sequence <TATAAA>) or CpG islands, a region rich in CG dinucleotides [146] as their core domains. In *Pyrocystis lunula* luciferase (*lcf*) genes, a GC box consensus sequence <GGGCGG> is present, but its location is further upstream than the usual position of -110 (numbered relative to the transcriptional start site at +1) as found in many eukaryotes [147]. Furthermore, a GC-rich motif <C(G/C)GCCC> was also found within the upstream region of *P. lunula lcf A* and *L. polyedrum lcf* and *lbp* genes, but their position were not fixed. This GC-rich motif was first reported in the upstream region of the *Peridinium bipes* ferredoxin gene [148]. However, the role of this motif in gene expression has still not been established. Both TATA-box or CpG island type promoters may include additional sequence elements such as the GC-box <GGGCGG>, the CAAT-box <CCAAT>, and the INR box <(C/T)(C/T)AN(T/A)(C/T)(C/T)> at which transcription is initiated. Interestingly, the TATA box is quite conserved in eukaryotes and is also found in protists as diverse as amoebas (*Acanthamoeba*), slime molds (*Dictyostelium*), ciliates (*Histriculus cavicola*), and apicomplexa (*Plasmodium*) [149-154]. On the other hand, members of the phylum Parabasalia use their own specific promoter element instead of the canonical TATA box [155-157].

Proper understanding of gene organization and structure is required to describe transcription in dinoflagellates. For example, *L. polyedrum* has multiple copies of peridinin-chlorophyll *a*-binding protein (*pcp*), Luciferin binding protein (*lbp*) and Luciferase (*lcf*) genes arranged in long tandem repeats [158-161]. PCR with *Pyrocystis lunula* genomic DNA revealed that, among *lcf A*, *lcf B* and *lcf C* isoforms, two, *lcf A* and *B* are in tandem repeat. To test if this is a general character of dinoflagellates, PCR was used with primers directed away from one another in *Amphidinium carterae* [162]. PCR using genomic DNA as a template was expected to produce a band if the genes were found as a tandem repeat, and this strategy revealed that 17 out of the 47 genes tested did indeed have a tandem repeat structure.

However, in *L. polyedrum* the sequence of the intergenic region between *lcf* and *pcp* coding sequences did not contain any known promoter elements. The only common feature between the two was a conserved 13 nucleotide sequence, CGTGAACGCAGTG, proposed as a dinoflagellate specific promoter sequence [158] but no further work has been published to firmly establish this result. Moreover, this sequence is not conserved among different dinoflagellate species as it is absent in the intergenic region between *P. lunula lcf A* and *lcf B* genes [163].

The lack of identifiable sequence elements in the intergenic spacers has led to the suggestion that tandem gene repeats may form a polycistronic transcript, in a manner similar to the *Trypanosoma* gene structure [164]. The trypanosomes transcribe long polycistronic transcripts from a single promoter containing genes coding for different gene products, and the primary transcript is then processed into mature mRNAs by *trans* splicing of the SL leader at the 5' end and by polyadenylation at the 3' end. If true for dinoflagellates, one possibility would place a promoter upstream of each tandem array, thus explaining the lack of recognizable promoter sequences in the intergenic regions. However, the consequences of this hypothesis include the predictions that the intergenic spacer region should be abundant in the transcribed RNAs, and that sequence differences at a particular position between copies in low copy number arrays should be detected in the mature transcripts at a frequency that is higher than those detected for high copy number genes. In a recent transcriptomic study which addressed this issue, none of these predictions were validated experimentally [165].

Eukaryotic and prokaryotic transcription also differ in that three different RNA polymerases (RNAP) are used for the former while only one is used for the latter. The three eukaryotic enzymes have specialized functions, with RNAP I transcribing most ribosomal RNA (rRNA), RNAP II transcribing protein-coding messengers (mRNA), small nuclear RNAs (snRNA) and micro RNA (miRNA), and RNAP III synthesizing transfer RNAs (tRNA) and the 5S rRNA. An assessment of the activity of RNA polymerase in the dinoflagellate *Cryptothecodinium cohnii*, carried out with radiolabeled UTP, revealed that considerable amounts of RNA polymerase activity remained even after inhibition by α -

amanitin, a potent inhibitor of RNAP II. This thus confirmed the presence of multiple forms of DNA dependent RNA polymerase as in other eukaryotes [166]. Curiously, this research also noted a peculiar inhibition of polymerase activity by Mn^{+2} , instead of the activation of these enzymes seen in other eukaryotes. It was suggested that dinoflagellate RNAP II activity might differ slightly from the other eukaryotic RNAP II enzymes [166], perhaps analogous to the unusual form of RNAP II found in some trypanosomes [167]. However, the transcriptome of *L. polyedrum* contains a complete set of core and common elements of all the three eukaryotic RNAPs in the dinoflagellate. Furthermore, the specific elements absent from the transcriptome were also missing in other members of the Alveolata (Figure 1.2.3.). It seems that the alveolates in general can assemble functional RNAPs with a reduced number of components as compared to higher eukaryotes, and there is nothing unique to the dinoflagellates in this part of the transcriptional machinery.

1.2.3.2. Basal/General Transcription factors

In addition to RNAP II, an *in vitro* reconstitution of a functional eukaryotic transcriptional apparatus requires a suite of other basal/general transcriptional factors (TF) [168]. Six multi-subunit complexes, termed TFIIA, TFIIB, TFIID, TFIIIE, TFIIF and TFIIH, appear to be among the most important [169-173]. The first step of promoter recognition is performed by TFIID, constituted from the TATA binding protein (TBP) and at least 14 TBP-associated factors (TAFs) [174, 175]. TBP binding is considered to be the rate-limiting step in the transcription process [176], although TBP can have relatives, such as TBP-related factors (TRF), which also activate transcription from the same RNAP II promoters that are activated by TBP [177, 178]. These TRFs have been found in diverse animals, including fruit fly, nematode, frog, zebrafish, chick, mouse, and human [177]. Interestingly, *C. cohnii* has been shown to contain a TBP-like factor (TLF), clearly homologous to TBP yet lacking four phenylalanines known to interact with the TATA box. This TLF is unique to dinoflagellates (Figure 1.2.4.) and has a strong affinity for a <TTTT> sequence instead of the consensus TATA-box sequence [179]. Unfortunately, the upstream regions from 6 different genes of two different dinoflagellates did not contain a TTTT element [179]. This suggests a unique

promoter recognition mechanism for at least these genes, in keeping with the unusual structure of the chromatin of these organisms.

The *L. polyedrum* transcriptome contains two TLF isoforms similar to the TLF found in *C. cohnii* and, somewhat surprisingly, no TBP at all [165]. The phylogenetic relationship between the consensus TBP and the TLF, found uniquely in the dinoflagellates, clearly indicates the early divergence of TLF from TBP as well as the presence of two distinct TLF clades within the dinoflagellates (Figure 1.2.4.). In agreement with this lack of TBP, it is perhaps not surprising that *L. polyedrum* also lacks most other TAFs, although the closely related *Alexandrium* expresses two proteins with DNA helicase activity, RuvB-like1 and RuvB-like2 [180]. RuvB-like proteins have been shown to co-purify with the human RNA polymerase holoenzyme complex and found to be an extremely important element required for growth [181], suggesting they may also play a role in the dinoflagellates. In particular, *L. polyedrum* lacks any TFIIA, TFIIB, TFIIE or TFIIF components, and only 3 out of the ten expected TFIIH components are found (the E-value cut-off for the BLAST is e^{-25}). It must be noted, however, that ciliate, apicomplexan and diatom genomes contain a single TBP and also lack the TAFs and TFs missing in *L. polyedrum* [165]. A figurative representation of basal TF status in different eukaryotes (Figure 1.2.5.) indicates that the poor conservation of TAFs and other basal TFs in *L. polyedrum* is commensurate with the other related eukaryotes. These properties thus seem more likely to be due to a reduced dependence on these TFs throughout the Alveolata than to the unusual nature of the dinoflagellate chromatin.

1.2.3.3. DNA binding proteins

Histones are the most abundant and conserved class of basic proteins in the DNA-binding protein class of eukaryotes, and can profoundly affect transcription rates by their ability to alter the degree of chromatin condensation. The classic nucleosome structure, observed microscopically as “beads on a string”, forms when 146 bp of DNA wraps 1.65 times around the histone octamer (dimers of each of the four core histone proteins H2A, H2B, H3, and H4)

[104, 182]. A fifth protein, histone H1, binds to the linker DNA between nucleosomes to induce an even higher structural order to the chromatin [183]. Dinoflagellates have long been thought to lack histone proteins, and there is considerable biochemical evidence to support this view [184]. Dinoflagellate protein extracts do not show the typical pattern of histones after polyacrylamide gel electrophoresis [185, 186] and no nucleosomes are visible in dinoflagellate DNA spreads observed under a microscope [187, 188]. However, the presence of all the core histone sequences in the transcriptome of two dinoflagellate species, the high sequence conservation of these sequences compared to other eukaryotic histones, and the presence of a wide range of histone modifying enzymes in the *L. polyedrum* transcriptome all suggest that histone proteins are indeed expressed [102, 189], albeit at levels still undetectable by antibody or MS analysis [189]. The only other eukaryotic cells lacking histones are sperm nuclei, which instead employ arginine-rich proteins called protamines to stabilize their DNA structure [190, 191]. No protamines are found in the *L. polyedrum* transcriptome.

The total amount of basic proteins in dinoflagellate nuclei (basic protein to DNA ratio of 1:10 [192]) is much lower than generally found in eukaryotes (1:1 ratio [193]) and prokaryotes (1:1.75 ratio [194]) and appears to date to include two different basic protein types. One, a group of histone-like proteins (HLPs) [185], were first found by electrophoretic analysis of acid soluble nuclear proteins in the dinoflagellate *C. cohnii* and later renamed HCc 1-4 [112, 195]. Blast homology search with *C. cohnii* HLP revealed that *L. polyedrum* also has an HLP, which was named HLP [115], and this protein was shown to have sequence specific DNA binding activity and be subject to post-translational modifications suggesting that its activity might be regulated *in vivo* [115]. A second basic protein called DVNP (dinoflagellate/viral nucleoprotein), recently found in studies of the basal dinoflagellate, *Hematodinium*, can bind DNA as efficiently as histones and can also be post-translationally modified [196]. DVNP is found only in dinoflagellates, including the early diverging lineage *Hematodinium*, as well as in a family of large algal virus, the Phycodnaviridae. However, DVNP is not found in the basal dinoflagellate lineage *Perkinsus*, which has instead the typical eukaryotic chromatin with all core histone proteins and DNA arranged into nucleosomes [197]. The gain in DVNP thus occurred at some time following divergence of

Hematodinium and the main dinoflagellate lineages from *Perkinsus* and thus appears to coincide with the acquisition of the unusual core dinoflagellate nuclear morphology. In addition, a substantial proportion of the DNA appears to consist of repeated sequences, and it is possible that this may contribute to genome organization [198].

The nuclear matrix is a network of fibers in the nucleus that also plays a key role in the functional and structural organization of the chromatin. Electron microscopy studies of nuclear matrices in the dinoflagellate *Amphidinium carterae*, produced *in situ* by microencapsulation in agarose and sequential extraction coupled with the Immunoblotting, revealed the presence of two matrix proteins (lamins and topoisomerase II) similar to what is found in higher eukaryotes [199]. The lamins are architectural proteins, a class of intermediate filaments that line the inside of the metazoan nuclear envelope and act as a scaffold to which proteins and chromatin bind [200]. They have a wide range of nuclear functions such as higher-order genome organization, chromatin regulation, transcription, DNA replication and repair [201, 202]. Thus, although the dinoflagellate chromatin is arranged differently from other eukaryotes, its nuclear matrix is conserved, perhaps indicative of an ancient evolutionary trait required for nuclear structure.

In pursuit of sequence-specific DNA binding proteins (as opposed to basal or general TFs), a dinoflagellate nuclear associated protein (Dinap1) was found in *C. cohnii*. Dinap1 does not have any known homologues but does contain two zinc finger domains (known to be present in many transcriptional factors) and two WW domains (known to interact with proline-rich domains) [203]. An interaction study using the Dinap1 WW domains identified five proline-rich Dinap1-interacting proteins (Dip) [204], and screening of a *C. cohnii* cDNA library with a tagged Dip1 retrieved not only the expected Dinap1 but also other interactants, named DAP (Dip1-associated proteins) [204]. Dinap1, Dip1 and DAP were all found in the nucleus and all have the same pattern of protein expression. Unfortunately, none of the above-described proteins interacted with DNA directly [204], although some as yet unidentified intermediate partners may be involved in DNA recognition. In addition to Dinap1, a homologue of the Tubby-like protein (TUBL) [180], a group of membrane-

tethered transcription factors involved in the signaling pathway [205] has been found in *Alexandrium*, although this protein has not been fully characterized [180].

Gene specific transcription factors (TFs) are one of the largest family of proteins, accounting for ~ 4% of the genome in yeast or ~ 8% of the genome in plants and mammals [206]. In contrast, proteins with a DNA binding domain account for only 0.15-0.3% of the total transcripts in each of two different dinoflagellates, *Lingulodinium* and *Symbiodinium* [102, 165]. Furthermore, in both species, roughly two-thirds of the TFs are represented by a single group, the Cold Shock Domain (CSD) containing proteins. The CSD is relatively uncommon in eukaryotes, and importantly, is more often implicated in posttranscriptional than transcriptional regulation [207]. Whether or not the dinoflagellate version of the CSD proteins will be shown to be *bone fide* DNA-binding proteins, and the reason for the preferential expansion of this domain in dinoflagellates, remains to be discovered. However, there is a caveat to assuming that dinoflagellates are bereft of most DNA binding domains based on gene sequence data. Apicomplexans were initially also thought to have a low number of DNA binding proteins, yet further research revealed the expansion of a unique family of transcription factors, ApiAP2, in these organisms [208]. An as yet unknown family of factors modulating transcription may remain to be discovered in the dinoflagellates.

1.2.3.4. Transcriptional regulation

Methylation of cytosine in the DNA is a well-studied epigenetic modification that plays an important role in several cellular processes such as retrotransposon silencing, genomic imprinting, X-chromosome inactivation, regulation of gene expression, and maintenance of epigenetic memory [209]. Cytosine methylation occurs at roughly 0.5-4% of cytosines in dinoflagellates [120, 210], and is dynamic as it has been shown to change with varying light conditions [211]. It is thus possible that cytosine methylation may structurally regulate the access of DNA to transcription. In addition to 5-methylcytosine (5-MeC), dinoflagellates possess a number of unusual base modifications such as 5-hydroxymethyluracil (5-HMeU) and N6-methyladenine (N6-MeA) [212]. 5-HMeU is formed in DNA as a product of oxidative attack on the methyl group of thymidine [213], and between 12-70% of the

thymidine in the DNA is found as 5-HMeU [121]. The significance of this modification in dinoflagellate DNA is still unclear.

The posttranslational modification of histones plays an important role in regulating gene expression in other eukaryotes, and deserves re-examination in dinoflagellates because of the recent discovery that conserved sequences for core histones and their regulatory enzymes appear in the transcriptomes [102, 189]. It is possible that very low levels of histones are associated with gene regulatory sites, much as the low levels of acetylated histone H3 are associated with initiation of polycistronic transcripts in kinetoplastids [214]. The role of HLPs in regulating gene expression is also unclear, although the sequence-specific DNA binding and their existence in several post-translationally modified forms may indicate an involvement in gene regulatory mechanisms [115]. HLP transcript abundance in dinoflagellates appears to be up-regulated during different phases of cell cycle and in response to nutrient availability, as exemplified by *Pyrocystis lunula* where HLP transcripts peaked during the S-phase [215] and *Alexandrium fundyense*, where HLP transcripts were up-regulated during G1 phase [215]. However, unlike the higher eukaryotes whose histone mRNA levels increase during S-phase, no difference in histone mRNAs abundance was found during S-phase in *L. polyedrum* [189]. It will be interesting to examine the newly discovered DVNP [196] to see if transcriptional regulation accompanies DNA synthesis in the dinoflagellates.

Most organisms have evolved an ability to respond to environmental changes including biotic and abiotic stresses such as changes in light or temperature. The signaling pathways involve receptors that sense and transmit the information to regulatory molecules, and changes in gene expression are a frequently observed cellular response [216]. For example, in *Amphidinium carterae*, Northern blot hybridization revealed that transcript levels of two light harvesting proteins, peridinin chlorophyll *a* protein (PCP) and a major a/c-containing intrinsic light-harvesting proteins (LHC), were, respectively, 86- and 6-fold more abundant under low light conditions than under normal light conditions [211]. Interestingly, this increase in transcript levels coincided with a decrease in DNA cytosine methylation of CpG and CpNpG motifs present near or inside the coding regions of the two genes under low light

intensity, although *in vitro* experiments to link DNA demethylation with transcriptional activation were unsuccessful [211]. *Karenia brevis* may also have a transcriptional response to low light, as the abundance of 9.8 % of the 4269 unique genes in the microarray differed between day and night [217]. In addition to light, temperature is also an important signal, and has been implicated in the loss of cnidarian-dinoflagellate symbiosis, a phenomenon called coral bleaching. Temperature increases induce oxidative stress in *Symbiodinium bermudense* that result in increased levels of superoxide radicals and hydrogen peroxide [218], and this may be the primary reason for loss of the symbiont [219]. To check the regulation of expression of heat shock protein (*hsp*) genes in *Symbiodinium* residing inside its coral host *Acropora millepora*, qPCR was used with samples that were subjected to elevated temperatures rapidly or gradually [220]. Dinoflagellate *hsp70* transcript levels increased from 39% to 57% when temperature increased to 26°C (moderate) or 29°C (severe), although when cells were exposed to extreme heat stress *hsp70* transcript levels decreased by up to 70%. Curiously, *hsp90* transcript levels always decreased under heat stress and were independent of the speed of the temperature increase [220].

Oxidative stress is often able to induce a transcriptional response in organisms. In *L. polyedrum*, metal-induced oxidative stress resulted in sharp increases in the activity of the defense enzyme superoxide dismutase [221], with the increase in activity dependent on the type of metal, its exposure time and concentration [222, 223]. This same stress resulted in an increase in the chloroplastic Fe-SOD transcript level which accounted for the increased enzymatic activity, clearly demonstrating the transcriptional response [224]. Similarly, a microarray of 3500 genes from *P. lunula* revealed that 204 and 37 genes increased in abundance by 2- to 4-fold after treatment with 1 mM sodium nitrite or 0.5 mM paraquat, respectively [225]. The transcriptional response of the heat shock protein genes *hsp70* and *hsp90*, to elevated temperature, metal and endocrine disrupting chemicals, were tested in the dinoflagellate *Prorocentrum minimum*. RT-PCR results revealed that Hsp70 transcripts increased in response to each of these stresses, while Hsp90 transcript level increased only in response to temperature and metals [226, 227]. Lastly, 454 pyrosequencing in the basal dinoflagellate, *Oxyrrhis marina*, revealed 9 and 21 transcripts to be up- and down-regulated by saline stress, respectively [228]. However it is worth mentioning that transcript levels of

only 11 of these 30 genes varied by more than 2-fold, and among these latter, 10 were in the down-regulated class. Clearly, dinoflagellates respond to a variety of stress conditions.

The circadian (daily) clock is an endogenous timer that regulates daily rhythms in organisms from all walks of life [68, 229-231], and although the clock receives timing cues from light/dark cycles or temperature changes [232-235], it provides signals distinct from these environmental conditions since rhythms can be maintained under constant conditions. Circadian rhythms presumably make organisms more fit by allowing them to specialize for different tasks at different times of day, and, in many cases, the physiological rhythms regulated by the clock are mediated through changes in gene expression. Indeed, microarray studies showed that the number of circadian mRNAs varied from 5-20% in *Neurospora*, 10% in *Arabidopsis*, 5 - 10% in mice and 30-65% in the cyanobacteria *Synechococcus elongates* [236, 237]. In the dinoflagellate *P. lunula*, 3% of the genes on the microarray were found to exhibit changes in transcript abundance (between 2- and 2.5-fold) [238] while in *K. brevis* 0.7% of the genes varied in both light/dark and constant light (between 2 - and 7-fold) [217]. The fluorescence labeling of total RNAs and ³²P incorporation of ribosomal RNAs in the stationary phase cells of *L. polyedrum* under constant light followed by subsequent gel electrophoresis of the labeled RNAs showed circadian rhythmicity with maximum RNA abundance at CT18 [239], the time corresponding to the peak of S-phase in these species [240, 241]. However, when *L. polyedrum* cells were treated with Actinomycin D (ActD), a drug that inhibits DNA-dependent RNA synthesis, the bioluminescence and photosynthesis rhythms were unaffected for 30 hours or more depending on the dose of the treatment [242]. In contrast to the lack of effect using transcription inhibitors, treatment with translation inhibitor puromycin causes an immediate inhibition of the rhythms [242]. As ActD will also indirectly inhibit protein synthesis, when RNA levels have decayed sufficiently, it is possible that the eventual loss of the rhythms by ActD treatment was due to decreasing levels of RNA. Similar tests with high concentrations of other potent inhibitors of RNAP II, such as DRB (5, 6-dichloro-1-beta-D-ribofuranosylbenzimidazole) and α -amanitin confirmed no significant effect on growth, luminescence or rhythmicity in *L. polyedrum* cultures [243]. Indeed, all circadian changes of protein levels in *L. polyedrum* have so far proven to be regulated post-transcriptionally [71].

Nutrient availability is also an important environmental cue, and can result in the formation of algal blooms for some dinoflagellates. The nutrients most important for the blooms are nitrogen (N) and phosphorus (P), and thus the transcriptomic response of dinoflagellates to N- and P-deplete and -replete conditions has been of great interest. When *Karenia* grown in N-deplete and -replete conditions were compared, 1102 genes on a microarray chip of 11000 genes were found to be differentially expressed [244]. Among the up-regulated genes were found type III glutamine synthetases, nitrate/nitrite transporters, and an ammonium transporter, all known to function in the nitrogen uptake and assimilation pathway. The transcriptomic response to P-depletion was not so informative, although 12% of the array showed a different expression profile. Interestingly, N and P concentrations and growth stages have a strong impact on the toxin levels produced by *Alexandrium tamarense*, suggesting that expression of genes involved in these pathways may be responsive to nutrients [245]. Microarray experiment with 4298 sequences from *Alexandrium minutum* identified 87 genes that specifically responded to N or P limitation [246], while massively parallel signature sequencing (MPSS) in *A. tamarense* cultures showed only 2 and 12 out of a total of 40,029 signatures were uniquely expressed under N and P starvation, respectively [247].

The strain and growth stage of dinoflagellate cultures can also affect gene expression. In the microarray study of *A. tamarense* discussed above, 489 of the 4298 sequences examined were found to be differentially expressed when exponentially growing and stationary phase cultures were compared, a number even higher than the response induced by nutrient deprivation [246]. Here, proliferating cells showed a greater abundance of translation pathway gene transcripts and a lower abundance of transcripts from genes involved in intracellular signaling [246]. Similar studies in *A. catenella* revealed proliferating cells show over-expression of transcripts from several categories, including transcription and RNA processing, protein synthesis and translational regulation, cell division, transport related, photosynthesis and cellular metabolism [180]. In *Karenia brevis*, five time points representing different growth phases were selected for microarray analysis, and taken together, 21% of the 11,000 features examined had accumulated to different levels in logarithmic compared to stationary phase cells [248]. Interestingly, a comparison of *toxic*

and non-toxic strains of *A. minutum* has indicated a strain specific regulation of gene expression [249]. Using microarray chips with a cut-off value of 1.5 fold difference, 145 and 47 contigs were identified as up-regulated in either toxic or non-toxic strains, respectively. While one of the original goals was to identify toxin-related genes in *Alexandrium*, it is unclear how much reliance can be placed on this line of experiments as many toxin related genes could also have unknown and important metabolic functions in the dinoflagellates and thus be similarly regulated in both strains. This view is supported by the observation that a non-toxic strain of *Heterocapsa circularisquama* transcribes a substantial number of genes thought to be involved in toxin biosynthesis [250].

Gene expression in the dinoflagellates can also be influenced by biotic factors, as shown by a massively parallel signature sequencing MPSS comparison of *A. tamarense* grown axenically and in normal cultures [247]. From a total of 40,000 signatures, 307 were differentially expressed in the axenic cultures (39% up-regulated and 61% down-regulated). The association of bacteria with the dinoflagellates seems to affect the methionine-homocysteine cycle and photosynthesis, as these categories were enriched in the differentially expressed genes. However, it is likely that the most important biotic factors will be those related to symbioses. The first indication of symbiosis-specific gene expression in dinoflagellates was obtained from study of *Scrippsiella nutricula* with and without its radiolarian host *Thalassicola nucleata*. It was found that several genes in the dinoflagellate were differentially transcribed depending on symbiotic or free living growth [251]. The dinoflagellate–cnidarian symbiosis, vital for ocean reef ecology, also presents an excellent model for understanding the regulation of gene expression by biotic factors. In this context, a homologue of P-type H⁺-ATPase gene in *Symbiodinium* was shown to be expressed exclusively during the coral symbiosis [252]. Thermal stress, the primary cause of coral bleaching, induced different responses in the host and the symbiont, with the coral expression pattern much more important than the dinoflagellate symbiont [253].

It must be kept in mind that most of the gene regulation studies performed in dinoflagellates are expression-profiling experiments, which indicate mRNA levels and are thus determined by the balance between mRNA synthesis and degradation rates. Indeed,

mRNA degradation may play a major role in determining the transcript abundance [254]. So far, only half-lives of transcripts whose protein synthesis are regulated by the clock in the dinoflagellate *L. polyedrum* have been measured [243]. Thus, different mRNA levels obtained during the gene expression studies cannot be unambiguously ascribed to result from transcriptional regulation.

1.2.4. Splicing and the spliceosome

Several posttranscriptional modifications in the primary transcripts of eukaryotic cells are necessary to create a mature mRNA that can be efficiently translated, and of these, arguably the most important is the removal of the intervening sequences, or ‘introns’, that interrupt the coding sequence, or ‘exons’ [255-257]. Mammalian genomes are generally intron-rich, while in contrast, dinoflagellate genes contain very few or lack introns completely. For example, all the high copy number genes tested in *L. polyedrum*, such as *pcp*, *lbp* and *lcf*, lack introns [158, 159, 161]. However, in another bioluminescent dinoflagellate, *P. lunula*, a comparison of genomic and cDNA PCR products of the *lcf C* gene identified a 403 bp intron [163]. The form II Rubisco gene lacks introns in *Prorocentrum minimum* [258], yet contains six introns in *Symbiodinium* [259]. The saxitoxin pathway gene *sxtG* in *Alexandrium* was found to have one intron whose length varied from species to species, ranging from 260 to 750 bp. Sequencing of different *sxtG* introns showed >90% intraspecies identity and <80% interspecies identity, with no variation observed within a strain [260]. Analyses of *hsp90* sequences from the genomic DNA of 17 dinoflagellates reported introns in only three species (97 bp, 134 bp and 289 bp in *Peridinium willei*, *Polarella glacialis* and *Thecadinium yashimaense*, respectively) [261]. A more detailed test, carried out with 31 genes in *A. carterae*, showed that four genes (encoding polyketide synthase, translation initiation factor 3 subunit 8, small nuclear ribonuclear protein and *psbO*) had 6 or more introns, similar to other eukaryotes, another 11 genes had less than 5 introns, and the rest no introns at all [162]. This study also correlated highly expressed genes with a very low intron density and a tandem gene arrangement in the genome.

The cellular mechanism that joins exons together by excising the introns is called splicing [256, 257]. As expected, splicing must be extremely accurate, as even a single nucleotide

frame shift could result in a nonsense mutation or a truncated protein. All introns in the nuclear-encoded pre-mRNAs are delimited by splice sites, which are critical sequences specifying the extremities and eukaryotic introns are generally bounded by the conserved dinucleotides GU and AG at their 5' and 3' ends respectively. Another important sequence, the branch point, is usually located between 18 to 40 nucleotides upstream from the 3' end of the intron, but except for a mandatory adenine which is ligated to the 5' end of the intron during the splicing reaction, its sequence is only loosely conserved. Interestingly, the dinoflagellate introns typically lack the usual GU-AG splice sites, as exemplified by the AT-TC intron found in *lcf C* of *P. lunula* [163], the G(C/A)-AG introns in *Symbiodinium rubisco* [259] and the AG-AG intron in the *Symbiodinium sxtG* [260]. Some of these novel splice sites have been shown to function in other eukaryotes, for example the introns with GC – AG boundaries described in animal and plant genomes [262].

The splice sites in pre-mRNA introns are recognized by base pairing to short RNA molecules (U1, U2, U4, U5 and U6) termed small nuclear RNAs (snRNA), each of which is bound to a complex of proteins to form small nuclear ribonucleoproteins (snRNPs). These five snRNPs, together with numerous non-snRNP proteins, constitute the spliceosome, a dynamic complex that forms and reforms repeatedly to process pre-mRNAs to mature transcripts [263]. Many of the protein components are highly conserved between mammals and dinoflagellates, as evidenced by the observation that autoimmune antibodies recognizing the so-called Smith antigen (Sm protein) present in all five human snRNP complexes were found to recognize four of the *C. cohnii* snRNPs [264]. In addition, the *L. polyedrum* and *Symbiodinium* transcriptomes contain sequences with significant homology to 70% and 85% of the splicing components, respectively [102, 165]. A high degree of sequence conservation was also noticed between the dinoflagellate and mammalian U2, U5 and U6 RNAs and, as in higher eukaryotes, the dinoflagellate Sm tends to protect an AUn region in the snRNAs [264]. Furthermore, the snRNAs of dinoflagellates have a modified 5' trimethylguanosine (TMG) cap, as do snRNAs of other eukaryotes [264]. Intriguingly, the spatial organization of the splicing process in the nucleus also appears similar in dinoflagellates and other eukaryotes. Several phylogenetically different species, including *Prorocentrum micans*, *Alexandrium fundyense*, *Akashiwo sanguinea*, and *Amphidinium carterae* were examined

microscopically after immunolabelling with antibodies directed against Sm proteins, DNA and p105-PANA (proliferation associated nuclear antigen) in conjunction with cytochemical staining for RNA, phosphorylated proteins and DNA [265]. These studies revealed a cross-reaction of the anti-Sm with eukaryotic-like perichromosomal granules, structures enriched in splicing factors that are actively involved in splicing, as well as Cajal-like bodies, nuclear regions thought to be involved in the modification and assembly of snRNPs. However, it must be noted that the anti-Sm labeling on Western blots revealed cross-reaction with proteins other than those of the expected molecular weight [265] raising the possibility that atypical Sm antigens may be present in the dinoflagellates.

Despite the paucity of cis-splicing events in dinoflagellates, trans-splicing is now known to be pervasive [266]. In this, dinoflagellates are similar to the kinetoplastid *Trypanosoma brucei*, where mRNAs were found to contain a consensus sequence of 39 nucleotides (nt) at their 5' ends. This sequence, termed a spliced leader (SL) sequence [267], is added from a separate SL-donor RNA (an SL RNA) in a process called trans splicing to all trypanosome mRNAs [268]. Since this initial discovery, many organisms including cnidarians, ctenophores, flatworms, nematodes, crustaceans, Euglena and now dinoflagellates have also been shown to use SL *trans*-splicing [269-271]. The length of the SL exon varies in different species, from 16 nt in *Ciona intestinalis* [272] to 51 nt in *Stylochus zebra* [273], and in dinoflagellates, the SL leader is a 22 nt sequence 5'-DCCGUAGCCAUUUUGGCUCAAG-3' (D = U, A, or G) [266]. The discovery of the dinoflagellate SL has provided an enormous boost to the study of dinoflagellate molecular biology, in part because full-length sequences of dinoflagellate cDNAs can now be readily retrieved, but more importantly, because dinoflagellate sequences can now be isolated from complex mixtures such as RNA extracted from environmental samples or from organisms in symbiosis [274]. The dinoflagellate SL sequence is derived from SL RNAs of 50–60 nt and contains an Sm binding motif (AUUUUGG) in the exon, unlike all other SL RNAs where this conserved sequence is found in the intron [266]. SL trans-splicing is absent in organelle-encoded transcripts, although a unique type of trans-splicing was recently found in the mitochondria of diverse dinoflagellates. The mitochondrial *cox3* gene is encoded in two pieces that are transcribed separately then trans-spliced to form a complete coding *cox3* mRNA [275]. SL trans-splicing

is evolutionarily ancient for the dinoflagellates, also being found in the basal lineage of dinoflagellates, as *Perkinsus marinus* has nuclear-encoded transcripts with either an SL identical to the core dinoflagellates (SL1) or a truncated 21 nt SL with either A or G as the starting nucleotide (SL2) [276]. The function of SL trans-splicing is not clear. It is unlikely to be involved in mRNA stability or translation, as there was no difference in translation efficiency or stability between *trans*-spliced and non-*trans*-spliced nematodes mRNAs [271]. It has been proposed that in conjunction with polyadenylation it functions in the production of mature monocistronic transcripts from polycistronic transcripts, and it is still possible that it defines the 5' end of transcripts even though polycistronic transcription now seems doubtful [165].

The paucity of introns, as well as the presence of multiple relict sequences related to the SL in the 5' ends of dinoflagellate genes isolated from genomic DNA, has led to the proposal of a mRNA recycling mechanism whereby mature mRNAs are inserted back into the genome through a recombination process [277]. This hypothesis still requires a more comprehensive enquiry in diverse dinoflagellates, but if true, may shed some light on the origin of the plethora of tandem array genes in dinoflagellates. It is also interesting that alternative splicing, a process by which cells can generate several proteins through permutation and combination of exons from a single pre-mRNA, has been discovered for cyclin transcripts in *Perkinsus marinus* [278]. Alternative splicing may have been lost after divergence from this basal lineage as to date, it has not yet been observed for other dinoflagellates.

1.2.5. RNA transport and mRNA surveillance pathways

Nuclear pore complexes (NPC) are enormous protein complexes, ranging from 50 MDa in yeast to 125 MDa in mammals, which are present within the nuclear envelope and mediate nucleo-cytoplasmic transport [279, 280]. Though small molecules under 40 kDa can passively diffuse through NPC, larger mRNA molecules require a more complex energy-dependent and signal-mediated process [281]. The nuclear export pathway has been well characterised in yeast and higher eukaryotes, but does not appear to be conserved in

apicomplexans, as many of the important components are either absent or unrecognizable by homology search algorithms [282]. To date, no description of this pathway has been made in any dinoflagellate, and we have thus analyzed the *L. polyedrum* transcriptome to try and retrieve the components expected for RNA transport. There are three general classes of proteins required, those forming the nuclear pore itself and those soluble in either the nucleus or the cytoplasm. Compared to the components found in other eukaryotes, the most marked difference between the alveolates and other organisms appears to lie in those components used for construction of the pore (Table 1.2.1.). Apart from the conserved integral membrane proteins termed Nups, thought to anchor the pores in the nuclear membrane, it seems that lower eukaryotes either manage to construct this large molecular complex with far fewer elements than are required in mammals, or alternatively, employ some unique and as yet unidentified constituents. It would evidently be of great interest to examine the structure of the pore using electron microscopy to ascertain if the pore retains the eightfold symmetrical structure normally found in higher eukaryotes. In addition to the NPC, a plethora of nuclear and cytoplasmic *trans* acting factors are also employed to mediate RNA processing and transport in mammals and higher eukaryotes. The nuclear components include factors common to the different types of RNA as well as other specific factors for processing and maturity that facilitate the nucleo-cytoplasmic transport [283], and these appear to be conserved in the dinoflagellates. In contrast, only a third of the mammalian and half of the plant cytoplasmic components involved in nuclear transport are conserved in *L. polyedrum* and other alveolates (Table 1.2.1.).

Eukaryotes also employ a multistep “quality control” or surveillance pathway to selectively degrade the damaged or mutated mRNAs as a protective mechanism against aberrant protein synthesis. This concerted procedure starts with mRNA capping during transcription within the nucleus, and ends in the cytoplasm with the degradation of abnormal mRNAs. There are three main pathways, the first being nonsense-mediated mRNA decay. In mammals, this pathway interprets stop codons found 50 or more nucleotides upstream from the last exon boundary to be premature stop codons, principally because normal stop codons are typically located in the last exon [284, 285] and this process uses factors involved in capping or 3’ end processing of the pre-mRNAs as well as a large complex of nuclear factors

comprising the exon-junction complex (EJC) as a scaffold [286]. These mRNAs are then degraded to block synthesis of truncated proteins that might act as dominant negative or gain-of-function mutants. Curiously, despite the conservation of many of the components, intron/exon boundaries are not required to fulfill the same role in invertebrates and yeast although the implication of the EJC is not well defined in these systems. Nonsense-mediated decay appears to be operative in dinoflagellates, as many of the generally conserved components are found (Table 1.2.2.), but the mechanisms used may be more similar to yeasts and insects as dinoflagellate genes have a generally low intron density. The second pathway, termed nonstop-mediated mRNA decay, is used to detect mRNA molecules lacking a stop codon. These transcripts pose a problem in that ribosomes translating into the poly A tail stall and have difficulty dissociating from the transcript, thus reducing the number of ribosomes available for general translation [287]. This mechanism requires both a release of the ribosome and a degradation of the mRNA, but the components required for this remain to be fully characterized. Lastly, recognition of stalled ribosomes may also be involved in what is termed no-go mRNA decay [288], where ribosomes stalled during translation, perhaps because of unusual secondary structure elements in the transcript, are also targeted for degradation [286]. In general, dinoflagellates and other alveolates have a very poor conservation of the nuclear factors required for RNA surveillance (27% as compared to mammals) although the conservation of cytoplasmic factors is better (67% as compared to the mammals) (Table 1.2.2.).

1.2.6. Conclusions and perspectives

Considerable progress has been made in the study of dinoflagellate transcription, fuelled in large part by the recent availability of low cost sequencing. We show here that most of the expected players in the transcriptional machinery are found in dinoflagellates, at least with respect to their counterparts among the Alveolata. The exception to this general rule is that the specific transcription factors seem in large part to be reduced in quantity and type in the dinoflagellates. Thus, while general transcription carries on much as expected for the eukaryotes, the specific targeting of genes for transcriptional control may differ as a result of

the unusual chromatin organisation in this class. Further studies will now be necessary to confirm the biochemical activities of some of the more interesting components identified from the massive influx of sequence information.

Table 1.2.1. mRNA transport components

Number of components involved in RNA transport found in the *L. polyedrum* transcriptome. Gene sequences for various Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were tabulated. The Alveolates are represented by *L. polyedrum* (Lp), *Plasmodium falciparum* (Pf) and *Tetrahymena thermophila* (Tt). A cutoff value of e^{-25} was used to assess the presence of components.

| | | Mammal | Plant | Alveolata | | | Diatom |
|-----------------|-------------------|--------|-------|-----------|----|----|--------|
| | | | | Lp | Pf | Tt | |
| Nucleus | | 11 | 10 | 6 | 9 | 7 | 8 |
| Central channel | Nuclear basket | 4 | 1 | 1 | 0 | 0 | 1 |
| | Symmetrical Nups | 11 | 9 | 2 | 1 | 4 | 6 |
| | Central channel | 3 | 3 | 0 | 0 | 0 | 1 |
| | Spoke complex | 5 | 5 | 0 | 0 | 0 | 1 |
| | Lumenal ring | 3 | 1 | 0 | 0 | 1 | 0 |
| | Cytoplasmic tails | 8 | 6 | 2 | 2 | 3 | 3 |
| Cytoplasm | | 53 | 37 | 17 | 17 | 17 | 24 |

Table 1.2.2. mRNA surveillance components

Number of components involved in mRNA surveillance found in the *L. Polyedrum* transcriptome. Gene sequences for various KEGG pathways were tabulated. The Aveolates are represented by *L. polyedrum* (Lp), *Plasmodium falciparum* (Pf) and *Tetrahymena thermophila* (Tt). A cutoff value of e^{-25} was used to assess the presence of components.

| | | Mammal | Plant | Alveolata | | | Diatom |
|-----------|-------------------------|--------|-------|-----------|----|----|--------|
| | | | | Lp | Pf | Tt | |
| Nucleus | Cap binding complex | 2 | 2 | 0 | 1 | 2 | 1 |
| | EJC | 15 | 11 | 5 | 4 | 4 | 5 |
| | 5' capping | 2 | 2 | 0 | 0 | 1 | 2 |
| | Pre-mRNA processing | 14 | 13 | 4 | 4 | 4 | 8 |
| Cytoplasm | Nonsense mediated decay | 12 | 9 | 7 | 6 | 5 | 6 |
| | No-go decay | 3 | 3 | 3 | 2 | 2 | 3 |

Figure 1.2.1. Superphylum Alveolata

The diagram shows the schematic representation of the phylogeny of the Superphylum Alveolata, which is marked by the presence of the cortical alveoli. Splice leader *trans*-splicing is a common feature in all the members of the dinoflagellate clade, while *Oxyrrhis* and the core dinoflagellates lack histones and have a dinokaryotic nucleus.

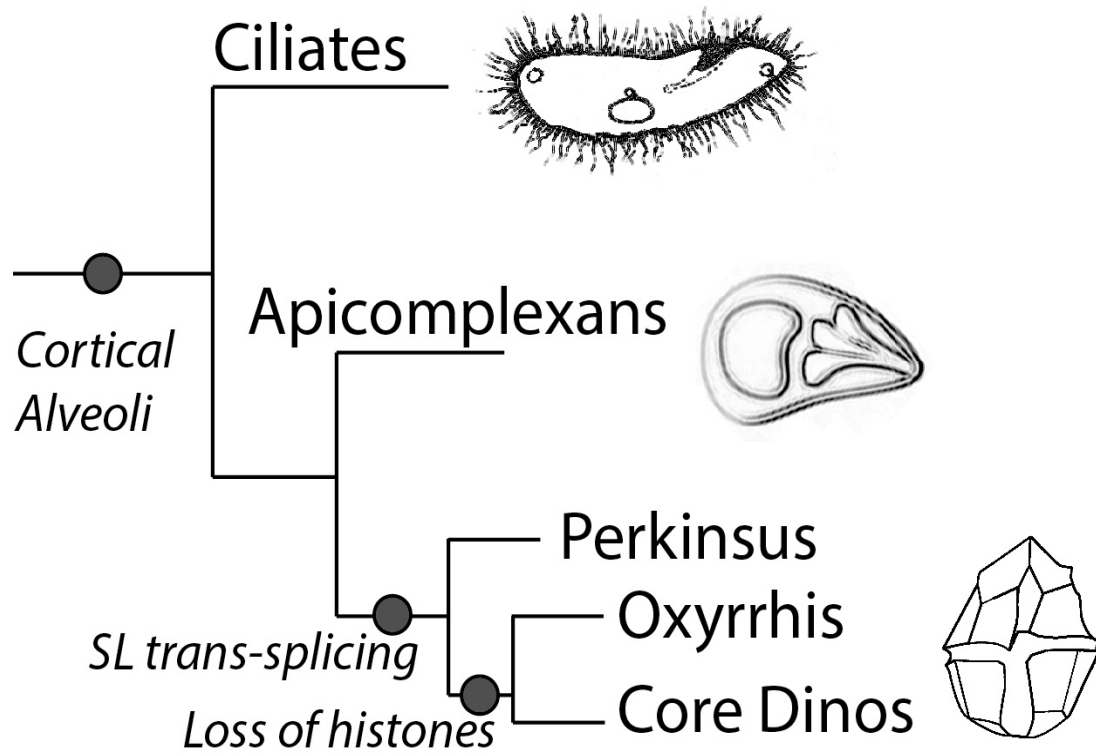


Figure 1.2.2. *Lingulodinium polyedrum* nuclear morphology

(a) Permanently condensed dinoflagellate chromosomes as visualized by fluorescence microscopy after DAPI. The C-shaped nucleus (n) is surrounded by the small punctate DNA staining of the multiple plastid genomes and lies under two larger spherical PAS bodies (p) at the apical end of the cell. (b) The nucleus viewed by transmission electron microscopy. The cross section shown lies near the back of the C-shaped nucleus (n) and shows chromosomes cut both in cross section (ovals) and longitudinally (cylinders), as well as plastids (p) and numerous diamond-shaped trichocysts. All scale bars are 10 μm .

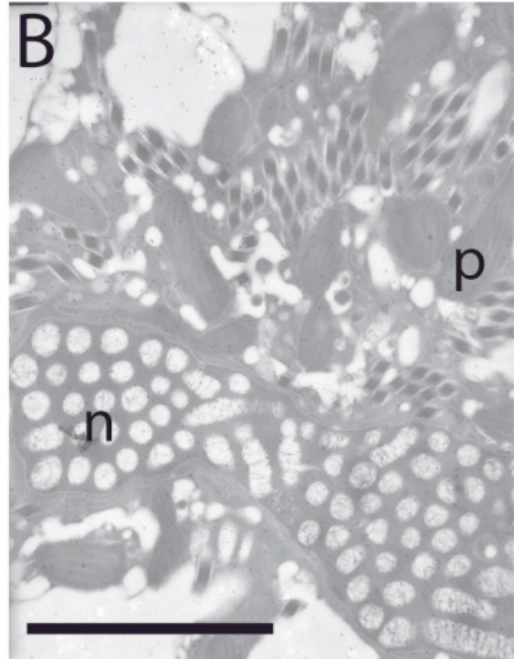
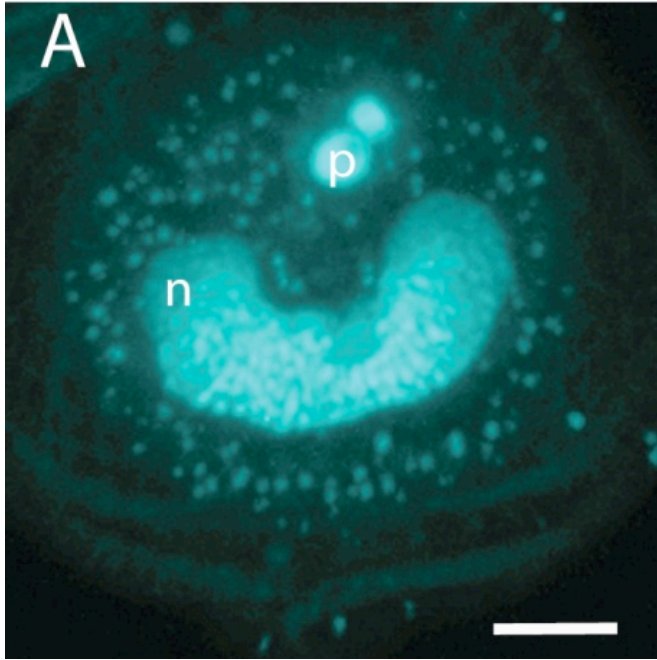


Figure 1.2.3. RNA polymerase components in *Lingulodinium*

The number of RNA polymerase components present over a wide phylogenetic range of organisms includes those considered to be core components (red), common components (yellow) and specific components (blue) of the RNAPI, II and III. Each bar represents an individual component. The representative sequences for the RNA polymerase I, II and III subunits were selected from an animal (*H. sapiens*), a plant (*A. thaliana*), a diatom (*T. pseudonana*), and two other alveolates (*T. thermophila* and *P. falciparum*) and uploaded and maintained as a local database in the Geneious software. Using tBLASTn and an expect E-value of $<e^{-25}$, the *Lingulodinium* transcriptome was scanned to obtain the homologues for the RNA polymerase subunits. For all other species the sequences were directly obtained from the KEGG specific pathway database by selecting the specific organism.

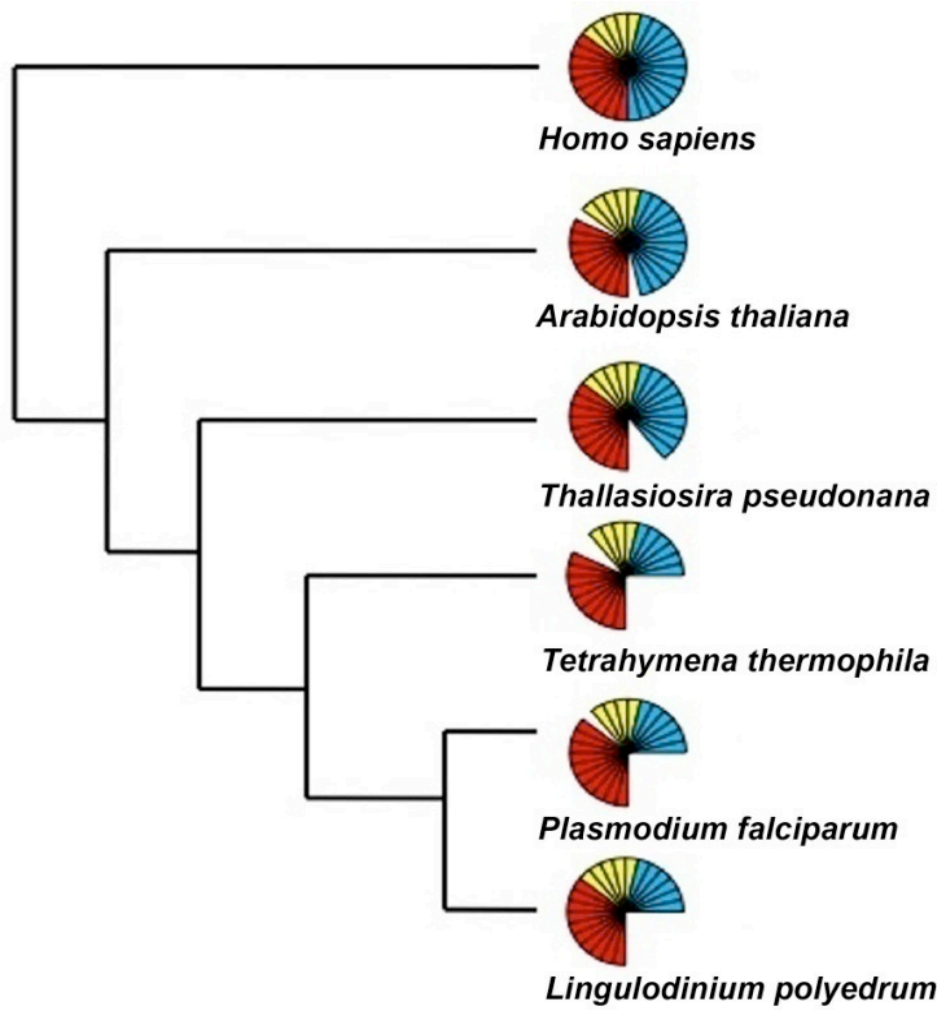


Figure 1.2.4. TBP phylogenetic classification

A simplified phylogeny of TBP and TBP-like proteins shows that the two TBP-like clades unique to dinoflagellates are distinct from all other TBP clades. The apicomplexan TBPs also form two clades, one from proteins in *Babesia* and *Toxoplasma* and the other from proteins in *Cryptosporidium* and *Plasmodium*. The protein sequences used include: **Animals** - *Homo sapiens* (CAG33057.1), *Mus musculus* (AAH50136.1), *Gallus gallus* (BAA20298.1), *Xenopus laevis* (NP_001084369.1), *Danio rerio* (AAQ07596.1), *Drosophila melanogaster* (AAA79092.1), *Strongylocentrotus purpuratus* (NP_999786.1); **Plants** - *Arabidopsis thaliana* (AEE75356.1), *Oryza sativa* (ABA99084.1), *Glycine max* (NP_001238202.1), *Zea mays* (NP_001105318.1); **Green Algae** - *Volvox carteri* (XP_002948268.1), *Chlamydomonas reinhardtii* (XP_001691004.1); **Diatoms** - *Phaeodactylum tricornutum* (XP_002186321.1), *Thalassiosira pseudonana* (XP_002293666.1); **Fungi** - *Neurospora crassa* (XP_960219.1), *Candida tropicalis* (XP_002548983.1), *Aspergillus nidulans* (XP_662580.1); **Alveolata** - *Cryptosporidium muris* (XP_002139943.1), *Cryptosporidium parvum* (AAR21861.1), *Tetrahymena thermophila* (EAR92317.1), *Toxoplasma gondii* (XP_002368492.1), *Ichthyophthirius multifiliis* (XP_004031283.1), *Babesia bovis* (XP_001610545.1), *Plasmodium vivax* (EDL43506.1), *Plasmodium falciparum* (XP_001351620.1), *Perkinsus marinus* (XP_002782410.1), (XP_002782409.1) and (XP_002782411.1), *Cryptocodium cohnii* (AAL24503.1), *Lingulodinium polyedrum* (JO752877.1) and (JO755256.1), *Symbiodinium* (kb8_c12831), (kb8_c27940), (mf105_rep_c7144), (mf105_rep_c14572) and (mf105_rep_c49191) [102]. For *L. polyedrum* and *Symbiodinium*, the translated sequences were aligned using MUSCLE, which is an in-built program in the tree construction software MEGA5 [289], used for the phylogenetic analysis here.

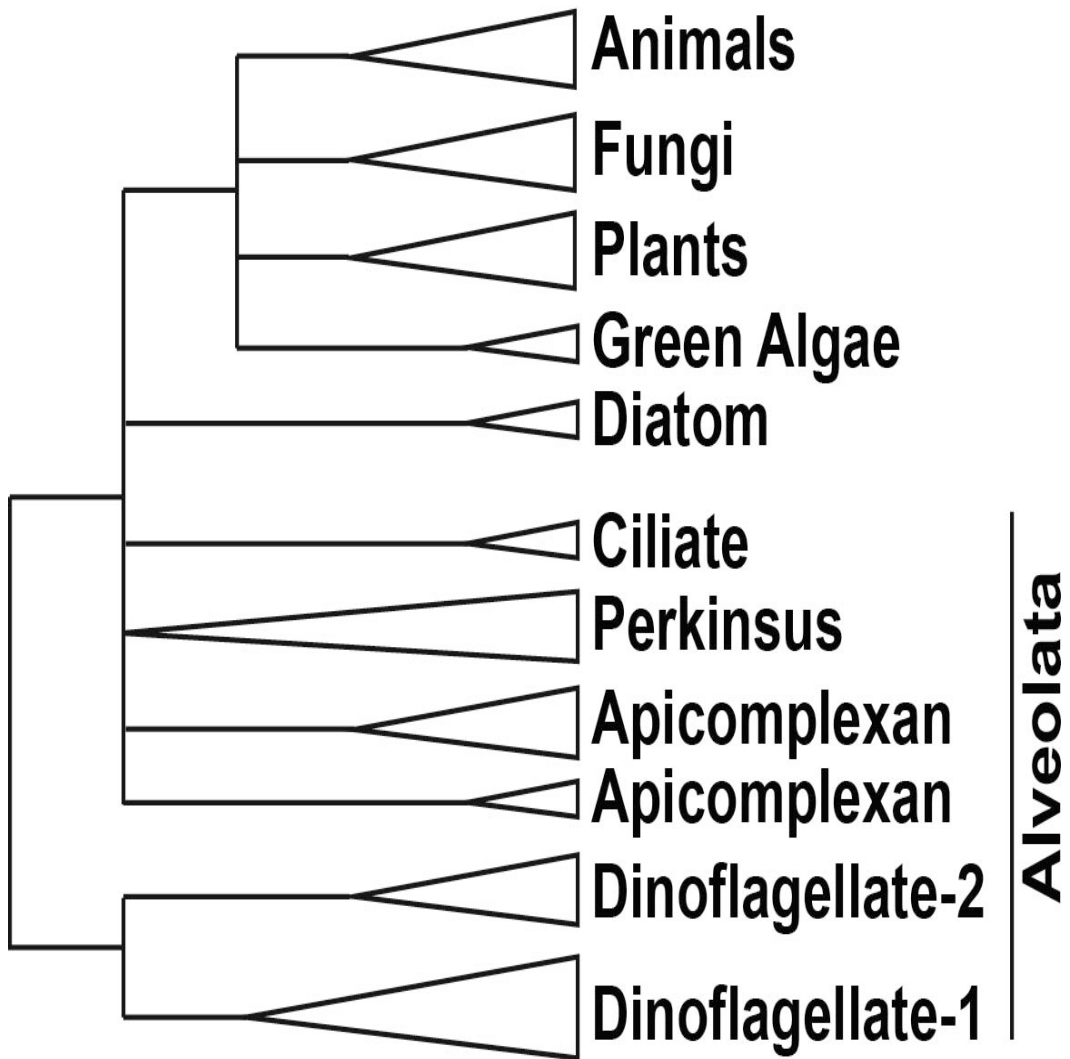
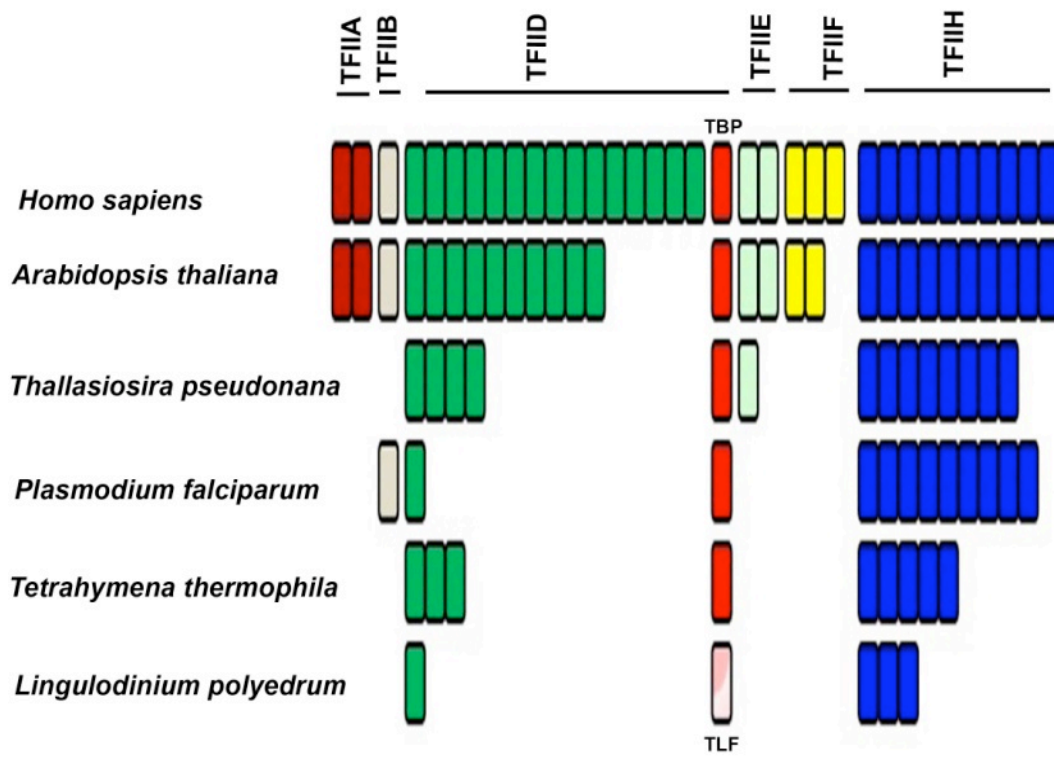


Figure 1.2.5. General transcription factors in dinoflagellates

Phylogenetic distribution of transcription factors associated with RNA-polymerase II shows a marked decrease in the number of TFII members among the apicomplexans. The dinoflagellates are the only group to use a TBP-like factor (TLF; pink) instead of TBP (red). Each bar represents a different component. A pool of basal transcription factor (BTF) protein sequences were selected from the 5 species then stored as a local database in Geneious. The *Lingulodinium* transcriptome was scanned using tBLASTn at an expect E-value of $<e^{-25}$, to obtain the homologues of the BTFs.



1.2.7. Acknowledgments

Research support from the National Science and Engineering Research Council (NSERC) of Canada to DM (Grant number 171382-03) is gratefully acknowledged.

1.3. Translation in Dinoflagellates

1.3.1. General Translation

Translation is a complex biological process where a sequence of ribonucleotides, termed a transcript or mRNA, is decoded to form a sequence of amino acids to form a protein. This consumes a considerable amount of the cell's energy budget- almost 5% of the human caloric intake and as much as 30 to 50% of the energy generated by rapidly growing *Escherichia coli* [290] - which underscores the enormous importance of this step. Translation is catalyzed by the ribosome, a molecular machine conserved across all kingdoms of life. A typical eukaryotic 80S ribosome is assembled in the nucleolus from four rRNA (called 18S, 28S, 5S and 5.8S molecules) and 79 ribosomal proteins (RPs), and is processed to small (40S) and large (60S) subunits before export to the cytoplasm. These two subunits must assemble around a mRNA molecule for translation to be initiated. Generally, several ribosome molecules are attached to a single eukaryotic mRNA, known as the polysomes or polyribosomes [291], which simultaneously code for many polypeptides from the same RNA. A high conservation of ribosome components is expected as they perform the common function in eukaryotic protein synthesis. Several researches demonstrated the conservation of ribosomal RNA (rRNA) secondary structure [292] as well as some ribosomal protein families that are highly conserved between evolutionarily divergent groups [293-295]. On the other hand poorly conserved RPs across species are also described that provides the necessary species specific diversity in the ribosome constitution [296]. Blast searches performed with known ribosomal protein sequences against the *Lingulodinium* transcriptome showed that the number of RPs in the *Lingulodinium* transcriptome is comparable to what is found in higher plant or mammalian genomes [165], and microscopic examination revealed that dinoflagellate nuclei have prominent nucleoli where rRNA is transcribed, suggesting a conservation of ribosome biogenesis mechanism [297]. However, when *Prorocentrum micans* 5.8S rRNA is compared with other eukaryotic and prokaryotic 5.8S rRNAs, a number of distinctive and dinoflagellates specific nucleotides have been observed. These distinctive nucleotides are found to be located in specific loops and could play a role in the ribosome organization [298]. Initiation represents the most highly regulated step in protein

synthesis, which also includes elongation and termination steps, and it is the initiation step that is often typically regulated to control the global rate of protein synthesis.

1.3.2. Translation factors

Eukaryotic ribosomes cannot bind to the mRNA sequences by themselves but require the help of at least 11 translation initiation factors (eIFs) [299], thus increasing the complexity as well as the number of potential regulatory steps [300]. Since eukaryotic transcripts have a monomethylated m⁷G (MMG) cap at the 5' end, translation normally requires the cap binding protein, eIF4E, and this binding is the usual rate-limiting step in initiating translation. The eIF4E binds to the cap in association with eIF4A, an RNA helicase, and eIF4G, a scaffold protein for assembly of the eIF4E and eIF4A. The identity of the 5' cap on dinoflagellate transcripts have not yet been fully described, although since all nuclear encoded transcripts have a trans-spliced 22 nucleotide splice leader (SL) sequence at the 5' end, the cap originating on the SL will be transferred to every mRNA transcribed in the nucleus. Preliminary data suggests the presence of a modified m⁷G cap on the SL-RNA although this needs to be established through further experimentation. As yet, there is no indication of modified cap structures in dinoflagellates, such as trimethylated guanosine (TMG) caps, as found in 70% of *C. elegans* transcripts that are *trans* spliced [301] or 2'O-methyl adenosine caps, found in many viral and cellular mRNAs [302]. Homology searches of the transcriptome of the dinoflagellate *Karlodinium* demonstrated that two families of cap binding proteins were present, eIF4E-1 and eIF4E-2 [303]. The qPCR analysis suggested that eIF4E-2 was more abundant than eIF4E-1, and affinity chromatography showed that only eIF4E-2 bound m⁷GTP-sepharose and neither isoform was able to bind TMG. In contrast, *C. elegans*, an organism with *trans*-spliced mRNAs have 5 different isoforms of eIF4E, some of which have different affinities for the MMG or TMG caps [301]. The ancestral dinoflagellate *Perkinsus*, encodes 8 eIF4Es, of which three, eIF4E-5, -6 and -7 group with eIF4E-2 from the *Karlodinium* [303].

Once bound to the cap, the eIF4 complex loads a 43S pre-initiation complex formed from eIF2-GTP, eIF3, the 40S ribosomal subunit and a Met-tRNA. This complex then starts

scanning through the 5' UTR for the start codon AUG. The recognition of the start codon AUG by the anticodon of the Met-tRNA is essential, but can be influenced by the flanking sequence. In mammals, an AUG in a consensus Kozak sequence $^A/G\text{CCAUGG}$, with conserved -3 (mostly A) and +4 (G) nucleotides positioned with respect to the +1 adenine of the start codon AUG, is favored over an AUG in another context [304]. The consensus sequences of yeasts (AAAAUGU) and plants ($A^A/_c\text{AAUGG}$) are similar but not identical [305-307]. Dinoflagellates seems to contain this consensus, as noted for the P-type proton pump in *Symbiodinium* for example [252]. However, in addition to scanning by the 43S pre-initiation complex, a second mode of translation initiation using an internal ribosome entry site (IRES) has been observed for a small group of eukaryotic mRNAs. This mechanism, which allows initiation within a mRNA is an exception from the normal 5' cap-dependent mode of translation. Comprehensive details of IRES have not been elucidated, however, it is known that they involve potential RNA motifs and specific trans-activating factors [308, 309]. There have been many efforts to identify common features in the cellular IRES but without much success, which has led to strong criticism of this mechanism [310-312]. As yet, no evidence of IRES has been reported in the dinoflagellates.

Poly (A) binding proteins (PABPs), along with cap binding proteins, are important elements whereby the integrity of an mRNA can be assessed before they are exported to the cytoplasm. In addition, PABPs can be involved in increasing mRNA stability and the rate of translation initiation to the point where, in mammals, PABPs are considered to be eIFs [313]. BLAST searches revealed multiple isoforms of the PABPs in the *L. polyedrum* transcriptome. Further characterization of eIF4E and PABP will be interesting, given the important roles they play in translation control. Apart from the eIFs, *L. polyedrum* also contains a plethora of eukaryotic translation elongation and termination factors [165]. Indeed, in sharp contrast to the paucity of transcription factors in dinoflagellates, eIFs are in general highly conserved.

We explored the two available transcriptome of *Lingulodinium* and *Symbiodinium* using blast homology searches for the known translational factors (Table 1.3.1.). It seems that dinoflagellates contain only one of the five subunits of eIF2B. The eIF2B complex facilitates the enhancement of non-enzymatic exchange of GTP for GDP on eIF2 [314] by

activating the eIF2•GDP to eIF2•GTP, and thus plays a major role in recycling of eIF2 [315] and eIF4B, which stimulates the activity of factors eIF-4A and eIF-4F [316], all other factors are conserved. However, in higher plants, the complete eIF2B protein complex has not yet been isolated, although homology searches identified all five eIF2B subunit.

1.3.3. Aminoacyl-tRNA synthetases

Another important component of the protein translation machinery are the Aminoacyl-tRNA synthetases (aaRSs), which catalyze two primary reactions: (1) activation of amino acids by adenylation and [317] transferring the activated amino acid to the 3' acceptor arm of the respective tRNAs [318]. Each of the 20 distinct aaRSs that is specific for particular amino acids have been thoroughly characterized [319]. Because of their ancient origin in the eukaryotic protein synthesis pathway, aaRSs can be used an excellent model for analysis of the selective forces that affect genome evolution [320]. As expected, the *Lingulodinium* aaRSs sequences that are found in the transcriptome are highly conserved, which is also evident from their high Expect (E) values (Table 1.3.2.).

1.3.4. Translational regulation

1.3.4.1. By protein factors

Although transcriptional regulation has received the lion's share of research attention, gene expression can also be modulated at a translational level and this is now known to play a critical role in development, differentiation, progression of the cell cycle, cell growth, and apoptosis [290]. Fine-tuning protein levels in the cells are often mediated via regulation of translation.

Many aspects of translational regulation are still not fully clear, although several novel and interesting characteristics have been elucidated. There are two general modes of translational control, the first of which one involves a global regulation of the translation rates. This is generally carried out through regulation of the global translational factors and the factors eIF4E [321], eIF2 [322] and PABP [323] have known critical roles. One well-studied example of this involves phosphorylation of the conserved serine residue in the alpha

subunit of eIF2 by a eIF2 α kinase. eIF2 is a GTP-binding factor involved in the formation of the 43S pre-initiation complex. This factor is charged with GTP in the functional complex, and GTP hydrolysis serves as a signal for recognition of the start codon by the Met-tRNA. To be reused, the GDP must be replaced with GTP, and this reaction is catalyzed by a guanine exchange factor called eIF2B [322]. Interestingly, the binding affinity of eIF2B to phosphorylated eIF2 α is much stronger than to non-phosphorylated eIF2 α , and as the amount of eIF2B in the cell is much less than eIF2 α , sequestration of the eIF2B by phosphorylated eIF2 α causes a generalized reduction in translation rates by blocking formation of the pre-initiation complex [324]. In higher plants, the dissociation constant of eIF2 for GDP is much lower than in yeast or mammals. Therefore it was assumed that eIF2 α phosphorylation to regulate general translation was not important in plants [325]. But in a recent study with *Arabidopsis* revealed the existence of a single *Arabidopsis* GCN2 kinase (in contrast to four in vertebrates) and proved its role in repression of global translation rates under different stress conditions [326]. Similarly, eIF4E-Binding Protein (4E-BP) family members are also known for global translation repression. They contain eIF4E recognition motifs and act as molecular mimics of eIF4G. The accumulation of hypo-phosphorylated (on threonine and serine residues) 4E-BP 1 sequesters the available eIF4E and either prevents its binding to the eIF4G or competitively displaces it [327, 328], thus inhibiting the association of 43S complex to the mRNA and repressing translation.

The second mode of translation control is one specific to a particular mRNA, and this type of control is normally mediated by structural features or regulatory sequences within the mRNA (termed *cis*-acting elements) that are recognized by *trans*-acting factors (either proteins or miRNAs capable of specifically binding the *cis*-acting elements). A specific control over translation can operate during formation of the pre-initiation complex, as exemplified by regulation of ferritin (an iron storage protein) synthesis by iron. Ferritin mRNA contains a sequence motif near its 5' end that can be recognized and specifically bound by protein factors called iron regulatory proteins, whose binding prevents assembly of the pre-initiation complex. The proteins remain bound as long as iron is absent, whereas the presence of iron prevents binding and thus allows synthesis of ferritin [329]. Specific control is also possible during the scanning of the 5'UTR by the preinitiation complex. One well-

studied example of this is the synthesis of GCN4, a transcription factor that mediates the response of yeast to amino acid starvation. The *gcn4* transcript has four small upstream open reading frames (uORFs) that are preferentially translated under conditions where amino acids are plentiful thus blocking translation of the GCN4 protein [330]. Interestingly, the mRNA encoding LBP in *Lingulodinium* has a uORF, and this uORF appears to influence which of two sizes of LBP produced by allowing selection of different translation start sites [331]. In general, there is a growing interest in translational control in many systems, and this is underscored by the lack of correlation in the levels of mRNAs and the proteins encoded by them [332].

A shutdown or reprogramming of protein synthesis can occur under different environmental conditions, including stresses such as nutrient depletion, temperature shock, DNA damage, and hypoxia. Interestingly, the dinoflagellate, *L. polyedrum* has been extensively studied as a model system for the programming of protein synthesis by an endogenous circadian (daily) clock, during which the synthesis rates of numerous proteins are regulated at a translational level [333]. Among the many different circadian rhythms studied, which include photosynthesis, bioluminescence, mitosis, nitrate reduction and swimming behavior, the bioluminescence mechanism has received particular attention. *Lingulodinium* bioluminescence is a luciferase-catalyzed oxidation reaction of a unique low molecular weight substrate (luciferin) [334]. In the cell, luciferin is normally bound to a luciferin binding protein (LBP) thought to protect the luciferin against uncatalyzed oxidation reactions [335]. Cellular bioluminescence is 40 to 60 times greater at night than day [334] and this correlates with rhythms in the abundance of both LBP [77], luciferin [336] and luciferase [337]. LBP levels change in the cell because the protein is rapidly and transiently synthesized for several hours during the early night phase and then preferentially degraded at the end of the night, by an as yet unknown mechanism [161]. Hence, LBP was the first example where timed synthesis of a protein was shown to be under the control of the circadian clock [77]. In addition, this control was found to occur at a translational level as although LBP synthesis rates changed, the amount of LBP mRNA remained constant over a 24-hour period. Sequence elements in the 5' and 3' UTR of the LBP mRNA, obtained from the complete cDNA sequence [161], were also investigated for potential binding of any

protein factors. Although no proteins were found that bind specifically to the 5' UTR, a protein was found that appeared to be a dimer capable of binding specifically to the LBP 3' UTR and exhibiting a circadian variation in its binding activity [338]. This binding activity correlated with the inhibition of LBP protein synthesis. However an attempt to purify and characterize this 3' UTR RNA binding protein (RBP) was not successful and subsequent studies were unable to confirm protein binding to the 3' UTR sequence either by electrophoretic mobility shift assay (EMSA), by cross-linking to LBP RNA *in vivo* or *in vitro*, by RNA binding to a cDNA library expressed in bacteria, or by yeast 3-hybrid assays [339]. Interestingly, however, when labeled LBP 3'-UTR from *Lingulodinium*, containing an unusual UG-repeat region, was incubated with *Chlamydomonas reinhardtii* protein extracts, three proteins (termed Chlamy1, 2 and 3) not only showed specific binding to the UG repeats but one of them (Chlamy1) was found to be clock controlled [340]. Subsequently, the daily change in *L. polyedrum* luciferase was also found to be mediated by translational control, as *lcf* mRNA levels were constant throughout the day [78]. The bioluminescence paradigm for clock control in *Lingulodinium*, in which clock mediated translational control was able to influence the levels of key players in the bioluminescence rhythm, led to the discovery of several other proteins whose synthesis was also shown to be under translational control. For example, *in vivo* pulse labeling experiments of early day and early night phase *Lingulodinium* cells with ³⁵S-methionine showed several proteins that were preferentially expressed in either early night or early day while no significant difference in labeled proteins were observed when the total RNA isolated from cells at the same two times were translated *in vitro* [341]. Among the proteins now known to be under circadian translational control are glyceraldehyde-3-phosphate dehydrogenase (GAPDH), ribulose biphosphate carboxylase/oxygenase (Rubisco) and peridinin chlorophyll *a* binding protein (PCP), all potentially involved in controlling circadian photosynthesis [342, 343]. Similarly, a superoxide dismutase isoform (Fe-SOD) and the TCA cycle enzyme NADP-dependent isocitrate dehydrogenase (NADP-ICDH) both show increased protein abundance during the day and may be involved in respiration [344]. Interestingly, rhythmic synthesis of proteins such as Rubisco or PCP does not cause a significant change in protein abundance, which might be due to low degradation rates compared to the rhythmic degradation seen for LBP and luciferase. Mass spectrometry studies in dinoflagellates have identified important groups

of proteins, whose abundance varied significantly as a function of growth phase, stress and light conditions [76, 345-347]. However without a comparison to mRNA abundance the extent of translational regulation cannot be predicted in these studies. It will clearly be of interest to compare proteomic and transcriptomic data in dinoflagellates to assess the extent of translation control.

Translational regulation has also been reported in other dinoflagellates. In *K. brevis*, a combination of quantitative PCR and antibody studies revealed that several S-phase genes, including proliferating cell nuclear antigen (PCNA), ribonucleotide reductase 2, replication factor C, and replication protein A, were all more abundant during S-phase even though their RNA levels were constant throughout the cell cycle [348]. A 2-fold variation in PCNA protein abundance without a change in mRNA levels was also found in *Pfiesteria piscicida* when exponential and the stationary growth phases were compared [349]. It must be noted, however, that constant mRNA levels do not point definitively to a translational control mechanism as degradation can also alter protein abundance. It is also possible that a combination of controlled synthesis and controlled degradation might be involved in fine tuning protein levels, as appears to be the case for LBP. However, post-transcriptional regulation of cell cycle genes appears to be the rule in *K. brevis*, where no significant cell cycle specific variation was noticed in their expression profiles of any of the 4600 unigenes on the microarray tested [217]. In contrast, higher eukaryotes, the S-phase specific activation of cell cycle gene expression is under transcriptional control by the transcription factor E2F [350].

The predominance of translational control in dinoflagellates does not correspond to the number of RNA binding proteins (RBP) present in the *Lingulodinium* transcriptome, as this class of protein is not over-represented when compared with other eukaryotes [165]. The use of BLAST searches using the major types of RNA binding domains, including RRM (RNA Recognition Module), RNA binding domain (RBD), DEAD box RNA helicase, PPR repeat and Pumilio (PUF) domains, indicates that they are also present in the dinoflagellates. However, this analysis is predicated on similarity of dinoflagellate sequences to known RNA binding proteins and will not reveal proteins not previously classified as such. As a case in

point, dinoflagellates contain a large number of cold shock domain proteins [102, 165], which are classified as DNA binding proteins. In point of fact, many of these proteins in other eukaryotes and bacteria are involved in the RNA metabolism pathway and in dinoflagellates they may actually be involved in binding RNA [351]. It is thus possible that the number and importance of RNA binding proteins in the dinoflagellate transcriptome has been underestimated.

1.3.4.2. By small RNAs

The phenomenon of RNA interference [352], first documented roughly 15 years ago [353], represents a mechanism whereby an injection with few molecules of double stranded RNA can induce an interference with gene expression. Small RNA molecules, termed small interfering RNAs (siRNAs) and microRNAs (miRNAs), are now considered potent elements for translation regulation in both the plant and animal kingdoms [354-357]. The siRNAs are either naturally formed or synthetic, are short double stranded RNA molecules, 20-25 base pair in length and with 2 nucleotide overhangs on the 3' ends of each strand. They can be formed by cleavage of longer double stranded RNA by an RNase III-like enzyme, named Dicer or Dicer-like (DCL). The short siRNA bind to an RNA binding protein of the Argonaute/piwi family, enzymatically denatured into single stranded guide RNA, which then as a part of a larger RNA-induced silencing complex (RISC), targets mRNA with a sequence complementary to the guide RNA for degradation. miRNAs are similar to siRNAs except that after transcription forms a pre-miRNA molecule, they are processed first to a roughly 60 base hairpin loop structure that is then exported to the cytoplasm where it is processed by Dicer. In many eukaryotes, an RNA-dependent RNA polymerase (RdRP), which is required for generating double stranded RNA from single stranded transcripts, is also involved in RNAi, typically acting to amplify the RNAi response [358].

The classical miRNA pathway, although widespread, is not ubiquitous, as it is absent in phylogenetically diverse organisms such as *Saccharomyces cerevisiae*, *Trypanosoma cruzi*, *Leishmania major* and *Cyanidioschyzon merolae* [359]. The pathway is also absent in the apicomplexan *Plasmodium*, a sister group to the dinoflagellates [360]. However, an abundance of antisense RNAs in the *Plasmodium* transcriptome suggests that an alternative

mechanism may replace the classical RNAi pathway [361]. Surprisingly, ciliates and the heterokont *Phytophthora*, belonging to the same kingdom as apicomplexans and dinoflagellates, contain all the three components required for miRNA biogenesis and formation of RISC complex [359]. Though other heterokonts such as the diatoms lack canonical dicer/DCL or RdRP genes, they do contain Argonautes as well as mi/siRNAs [362]. This suggests that an alternative to the dicer-mediated processing exists in these organisms, which can also generate these small RNAs. We have found DCL enzymes and Argonaute like proteins through BLAST searches in *Perkinsus marinus*, a common ancestor of dinoflagellate and apicomplexa, although we failed to detect any RdRp related proteins. Searches of the *Lingulodinium* and *Symbiodinium* transcriptomes found at least two of all three miRNA pathway candidates, a protein with RdRp domain as well as several piwi domain-containing proteins (four in *L. polyedrum* and two in *Symbiodinium*). A loosely conserved dicer-like enzyme, has also been found, which is consistent with what is seen in lower eukaryotes. But, unfortunately, the dicer-like transcript is incomplete at this time so its full domain structure is not known. The phylogenetic analysis of the RNA binding proteins found in the RISC complex shows two distinct classes, the first clade contains the plant AGO-like proteins along with other animals and algae, while the other clade is animal specific PIWI proteins with no representation from plants [359]. It will be interesting to find out the phylogenetic positioning of the Piwi-domain proteins of *Lingulodinium* and other dinoflagellates. Thus the presence of miRNAs in dinoflagellate is a possibility, which needs thorough investigation as it is an intriguing prospect for an organism known to exploit widespread translational control mechanisms in regulating gene expression.

It must be stressed that miRNAs can in some cases mediate translation rates without affecting mRNA levels and this is clearly of interest for circadian regulation of translation in *Lingulodinium* where no changes in RNA levels have yet been observed. Recent work suggests that the mechanism of translational repression is not substantially different from miRNA-mediated decay, and that repression may in fact be an intermediate in the degradative pathway [363, 364]. Unfortunately, no studies have addressed the possibility that miRNAs are found in dinoflagellates.

1.3.5. Posttranslational regulation of gene expression

Phosphorylation is by far the most important type of post-translational modification (PTM) of proteins found extensively in eukaryotes, although many other types are known. These include addition of carbohydrate groups (N-linked and O-linked glycosylation), small molecules (acetylation, amidation, hydroxylation, methylation), ubiquitin or ubiquitin-like modifiers (ubiquitination, SUMOylation), hydrophobic groups to increase membrane solubility (palmitoylation, myristoylation or prenylation), or proteolytic cleavage [221]. Indeed, while protein abundance has long been considered the most important step in regulating protein function, recent studies revealed the profound effect of PTMs in modulating the structure and function of proteins [365].

A recent census of PTMs shows that out of 72,430 experimentally characterized modifications, phosphorylation accounted for 49,090 of them [366]. In eukaryotes almost 30% of the cellular proteins can undergo phosphorylation [367, 368], though the significance of most of them are still uncharacterized. However, those phosphorylated proteins that are characterized is known to affect protein folding, enzyme activity, interactions between proteins, degradation rates and sub-cellular localization [367], which in turn can be used to regulate a wide variety of cellular activities such as intercellular communication, growth, proliferation, differentiation and apoptosis [369]. In addition to the prevalence of this PTM, phosphorylation is also amenable to analysis by high throughput methods, through a combination of phosphoprotein or phosphopeptide enrichment techniques and mass spectrometry (MS) sequencing. Unfortunately, the paucity of sequence databanks is an impediment to extensive studies of proteomics and phosphoproteomics in the dinoflagellates. To some extent this limitation has been alleviated by the recent developments in MALDI-TOF-TOF MS technology using a *de novo* sequencing strategy, and this approach has been used to characterize proteins with organisms having little or no genomic information available [370]. This strategy recently identified 158 unique proteins involved in different biological processes in *A. tamarense* indicating this powerful proteomics tool may allow characterization of proteins from unsequenced dinoflagellates.

Phosphorylation is also an integral part of circadian timekeeping in animal, plant, fungal and cyanobacterial models [81, 86, 371, 372]. In higher plants, as *Arabidopsis*, mRNA levels of several kinases and phosphatases are regulated by circadian clock [373, 374], and these in turn regulate different rhythmic phosphorylation/dephosphorylation events. In *Lingulodinium*, serine/threonine kinase and phosphatase inhibitors are known to affect the timing of the bioluminescence rhythm [93, 94], supporting a role for phosphorylation in dinoflagellate clock mechanism. It will be interesting to find out whether it is the kinases themselves or rather the kinase-substrates that are more important in the clock functioning in these organisms. As yet, we are unaware of the kinase repertoire in dinoflagellates. It will be thus interesting to find out the extent of kinase classes, their abundance and the range of their substrates at different times of the day/night cycle. Though some proteomics studies have been reported in dinoflagellates, phosphoprotein studies are very scarce. One such study showed the variation of PCNA abundance with cell cycle stages in *K. brevis* was also accompanied by a shift in its size agrees well with the fact that PCNA post-translational modifications, such as phosphorylation, which can control its activity [375]. 2-D PAGE phosphostaining and LC-MS/MS after phosphopeptide enrichment were used to ascertain the differential phosphoproteome of *L. polyedrum* at mid-day and mid-night respectively [376]. The phosphorylation intensity of 8 among the 45 phosphoproteins varied more than 2 fold between the two times studied. Three among them were RNA binding domain containing proteins [376], an interesting finding, as translational regulation is known to be predominant in these organisms. Though interference from a large number of acidic peptides seems to have prevented a sufficient enrichment of phosphopeptides, it certainly showed that detailed insight of the daily biochemical changes in *L. polyedrum* could be obtained through in depth profiling of its phosphoproteome.

The balance between synthesis and degradation rates of proteins determines its abundance in the cell at particular times. Unfortunately, although a body of knowledge is accumulating regarding control over protein synthesis, as yet no studies on protein degradation have been carried out in dinoflagellates. In other eukaryotic systems, a ubiquitous proteolysis mechanism involving addition of a small 76 amino acid protein (ubiquitin, Ub) to a target protein has been well-studied as a mechanism for controlling

access to the proteasome, an organelle specialized for protein degradation [377]. Addition of Ub involves E1 ubiquitin-activating enzymes, E2 ubiquitin-conjugating enzymes and E3 ubiquitin ligase enzymes, with the choice among many different E2-E3 complexes determining the specificity of the ubiquitylation [378]. The normal role of ubiquitylation is to allow entry of proteins to the proteasome, an organelle dedicated to protein degradation. In *Symbiodinium* strains 75-90% of the ubiquitin pathway elements are conserved [102]. Using the BLAST algorithm we describe the ubiquitin-proteasome components in the *Lingulodinium* transcriptome and compare them to close and distant relatives (Table 1.3.3.). In comparison to mammals *Lingulodinium* contains 40% of the E1, E2 and E3 components, which is equivalent to what is found in *Arabidopsis*.

The most well studied role for ubiquitylation is in mediating of protein degradation. However, the cellular consequences of ubiquitylation depend on the number and orientation of the ubiquitin moieties that are added to a target protein [379]. Target proteins can be mono-ubiquitylated or polyubiquitylated, and for those proteins with many Ub, they can be added to different amino acids in the target or to the same amino acid in a series or branched format. It is now becoming clear that ubiquitylation can be used to influence protein-protein interactions, most often to proteins containing ubiquitin-binding domains, and that this in turn can be used to regulate the activity and cellular localization of the target protein.

Table 1.3.1. Translation factors in dinoflagellates

This list shows the number of general translation factors found in the *L. polyedrum* transcriptome. A local database for the representative translation factors was prepared from Aveolates (*Plasmodium falciparum* and *Tetrahymena thermophila*), Diatoms, *Arabidopsis* and Human sequences in the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways database. *L. polyedrum* and *Symbiodinium* sp. transcriptomes were scanned using tBLASTn at an expect E-value of $<e^{-25}$ to obtain the homologues.

| Factor | <i>L. polyedrum</i> | <i>Symbiodinium</i> kb8 |
|---------------------------|---------------------|-------------------------|
| Initiation factors | | |
| eIF1 | 1.72E-16 | 2.45E-43 |
| eIF1A | 2.80E-73 | 1.10E-16 |
| eIF 2 subunit α | 2.20E-80 | 1.53E-69 |
| eIF 2 subunit β | 2.50E-58 | 3.31E-46 |
| eIF 2 subunit γ | 1.60E-175 | 0.00E+00 |
| eIF 2A | | 1.12E-19 |
| eIF 2B Alpha | None | None |
| eIF 2B Beta | None | None |
| eIF 2B Delta | 1.98E-46 | 9.14E-41 |
| eIF 2B Gamma | None | None |
| eIF2B Epsilon | None | None |
| eIF 2D | None | None |
| eIF 3 subunit 10 | 2.40E-66 | 2.80E-72 |
| eIF 3 subunit 8 | 6.70E-95 | 1.95E-68 |
| eIF 3 subunit B | 3.50E-101 | 1.13E-95 |

| | | |
|----------------------------|-----------|-----------|
| eIF 3 subunit D | 3.35E-76 | 1.89E-76 |
| eIF 4A | 6.90E-167 | 2.47E-165 |
| eIF 4B | None | None |
| eIF 4E | 1.10E-22 | 8.59E-24 |
| eIF 4G | 2.20E-25 | 5.80E-30 |
| eIF 5 | 1.68E-30 | 4.00E-30 |
| eIF 5A | 3.52E-19 | 4.60E-38 |
| eIF 5B | 3.00E-171 | 8.50E-165 |
| eIF 6 | 5.28E-91 | 1.04E-92 |
| PABP | 1.50E-120 | 4.02E-122 |
| Elongation factors | | |
| eEF 1alpha | 1.90E-131 | 1.10E-122 |
| eEF 2 | 2.33E-150 | 0.00E+00 |
| eEF 3 | 3.46E-91 | 1.13E-177 |
| EF-Tu | 1.50E-138 | 1.87E-140 |
| EF G | 5.60E-161 | 0.00E+00 |
| Termination factors | | |
| eRF 1 | 2.10E-165 | 1.60E-163 |
| eRF 3 | 9.60E-118 | 9.80E-118 |

Table 1.3.2. Aminoacyl-tRNA synthetases in dinoflagellates

This list shows the number of aminoacyl-tRNA synthetases found in the *L. polyedrum* transcriptome. Homologous sequences were obtained as described in the legend to Table 1.3.1.

| AA-tRNA synthetases | <i>Lingulodinium</i> | <i>Symbiodinium</i> kb8 |
|---|-----------------------------|--------------------------------|
| Glutamyl-tRNA synthetase | E-105 | E-114 |
| Glutaminyl-tRNA synthetase | E-93 | E-84 |
| Alanyl-tRNA synthetase | 0 | 0 |
| Aspartyl-tRNA synthetase | E-123 | E-113 |
| Asparaginyl-tRNA synthetase | E-141 | E-136 |
| Glycyl-tRNA synthetase | E-154 | E-151 |
| Threonyl-tRNA synthetase | 0 | 0 |
| Seryl-tRNA synthetase | E-157 (3) | not found |
| O-phosphoseryl-selenium-tRNA synthetase | E-72 | E-100 |
| Cysteinyl-tRNA synthetase | E-141 | E-101 |
| Methionyl-tRNA synthetase | E-85 | E-148 |
| Methionyl-tRNA Formyltransferase | E-67 (1) | E-71 |
| Valyl-tRNA synthetase | 0 | E-131 |
| Leucyl-tRNA synthetase | 0 | 0 |
| Isoleucyl-tRNA synthetase | 0 | 0 |
| Lysyl-tRNA synthetase | E-175 | E-180 |
| Arginyl-tRNA synthetase | E-132 | E-131 |
| Prolyl-tRNA synthetase | E-139 | E-140 |
| Histidyl-tRNA synthetase | E-42 | E-33 |

| | | |
|-------------------------------------|--------------|--------------|
| Phenylalanyl-tRNA synthetase | E-111 | E-111 |
| Tyrosyl-tRNA synthetase | E-121 | E-104 |
| Tryptophanyl-tRNA synthetase | E-148 | E-150 |

Table 1.3.3. Ubiquitin mediated proteolysis

This list provides the number of components involved in ubiquitin mediated proteolysis found in the *L. polyedrum* transcriptome. Gene sequences for various Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were tabulated. The Alveolates are represented by *L. polyedrum* (Lp), *Plasmodium falciparum* (Pf) and *Tetrahymena thermophila* (Tt). Plants and Diatoms are represented by *Arabidopsis* and *Thalassiosira pseudonana* respectively. tblastn was used to scan the *Lingulodinium* transcriptome with a cutoff value of e^{-25} to obtain the homologues.

| | | Human | Plants | Alveolata | | | Diatom |
|-----------------------|--------------------------------|-------|--------|----------------------|---------------------|-----------------------|--------|
| | | | | <i>P. falciparum</i> | <i>L. polyedrum</i> | <i>T. thermophila</i> | |
| E1 | | 4/4 | 4/4 | 2/4 | 3/4 | 3/4 | 4/4 |
| E2 | | 24/24 | 14/24 | 10/24 | 15/24 | 12/24 | 12/24 |
| E3 | HECT type | 15/15 | 4/15 | 1/15 | 5/15 | 2/15 | 4/15 |
| | U-Box type | 6/6 | 4/6 | 3/6 | 3/6 | 3/6 | 3/6 |
| | Single RING finger type | 20/20 | 5/20 | 1/20 | 3/20 | 1/20 | 3/20 |
| Cullin -RBX E3 | SCF complex | 4/4 | 4/4 | 2/4 | 3/4 | 2/4 | 3/4 |
| | ECV complex | 5/5 | 2/5 | 1/5 | 2/5 | 1/5 | 2/5 |
| | Cul3 complex | 3/3 | 2/3 | 1/3 | 2/3 | 2/3 | 2/3 |
| | Cul4 Complex | 4/4 | 4/4 | 1/4 | 2/4 | 1/4 | 4/4 |
| | ECS complex | 5/5 | 1/5 | 0/5 | 0/5 | 0/5 | 1/5 |
| | Cul7 complex | 4/4 | 2/4 | 2/4 | 2/4 | 2/4 | 2/4 |
| APC/C | | 14/16 | 11/16 | 2/16 | 6/16 | 11/16 | 10/16 |

CHAPTER 2 – PUBLICATION # 1

A full suite of histone and histone modifying genes are transcribed in the dinoflagellate *Lingulodinium*

Sougata Roy and David Morse (2012). Published in PLoS One. 7:e34340.

2.1. Abstract

Background: Dinoflagellates typically lack histones and nucleosomes are not observed in DNA spreads. However, recent studies have shown the presence of core histone mRNA sequences scattered among different dinoflagellate species. To date, the presence of all components required for manufacturing and modifying nucleosomes in a single dinoflagellate species has not been confirmed.

Methodology and Results: Analysis of a *Lingulodinium* transcriptome obtained by Illumina sequencing of mRNA shows several different copies of each of the four core histones as well as a suite of histone modifying enzymes and histone chaperone proteins. Phylogenetic analysis shows one of each *Lingulodinium* histone copies belong to the dinoflagellate clade while the second is more divergent and does not share a common ancestor. All histone mRNAs are in low abundance (roughly 25 times lower than higher plants) and transcript levels do not vary over the cell cycle. We also tested *Lingulodinium* extracts for histone proteins using immunoblotting and LC-MS/MS, but were unable to confirm histone expression at the protein level.

Conclusion: We show that all core histone sequences are present in the *Lingulodinium* transcriptome. The conservation of these sequences, even though histone protein accumulation remains below currently detectable levels, strongly suggests dinoflagellates possess histones.

Key Words: Histone, Dinoflagellate, Transcription, Translation

2.2. Introduction

Unlike typical eukaryotes, dinoflagellate chromatin is permanently organized into a cholesteric liquid crystal structure [116, 380], similar to structures observed in bacteria grown under stress conditions [381] or in sperm cell nuclei [190]. In the dinoflagellates, a combination of several factors may contribute to this structure, including a high concentration of divalent cations [382], a low ratio (1:10) of basic protein to DNA [383], and amounts of DNA that can range from 1.5 pg/cell (half that in a haploid human cell) in *Symbiodinium* [384] to roughly 200 pg/cell in *Lingulodinium* [385]. The unique chromatin structure in dinoflagellates is presumably a derived characteristic since nuclei in *Perkinsus*, a genus thought to be ancestral to the dinoflagellates [386], have a typical eukaryotic appearance [197].

An additional factor that is also likely to contribute to the unique structure of the dinoflagellate chromatin is the apparent lack of histones. This view is supported by biochemical evidence showing that protein extracts after gel electrophoresis lack the typical and distinctive pattern of histones [185, 186] as well as by microscopic observations showing that nucleosomes are not visible in DNA spreads [187, 387]. Instead of histones, dinoflagellates use histone-like proteins (HLPs) [111, 188]. HLPs of different dinoflagellates are similar but not identical [186], and have been shown to bind DNA and can be modified post-translationally [113, 115].

In general, DNA synthesis is coupled to histone protein synthesis for efficient assembly into nucleosomes. In plants and lower eukaryotes such as yeasts and ciliates, replication dependent histone mRNAs rely mainly on transcriptional regulation to affect histone accumulation in the S phase [388-390]. The N-terminal region of the histone proteins generally contains a nuclear localization signal (NLS) [391, 392] that binds to the nuclear import family of karyopherins with the help of Nucleosome Assembly Protein (NAP) [391-393]. Once inside the nucleus, the histones and DNA are assembled into nucleosomes by the help of NAP and other histone chaperone proteins [394, 395]. Certain residues in histone N-

terminal end undergo specific post-translational modifications such as acetylation, methylation, phosphorylation, ADP-ribosylation, ubiquitination, sumoylation and biotinylation [396]. Histone modification causes chromatin to reorganize and can result in epigenetic regulation of gene expression as well as affecting other DNA processes such as recombination, repair and replication [397].

A parsimonious explanation for the lack of nucleosomes and histones in dinoflagellate chromatin is that these organisms no longer contain or express histone genes. However, reports of histones H3 and H2A.X mRNA sequences in *Pyrocystis* and *Alexandrium* [225, 398] as well as by retrieval of all core histones and transcripts for two histone-modifying enzymes and a NAP from an environmental sample of dinoflagellates [274] cast considerable doubt on this idea. The environmental sample contains only dinoflagellate sequences, as their amplification exploited a splice leader (SL) sequence specifically trans-spliced to the 5' end of all nuclear encoded dinoflagellate mRNAs [266]. However, this study could not determine if any one species of dinoflagellate contained the complete set of histones or if the core histones were scattered among many different species and thus unlikely to be functional.

We undertook the present study because a transcriptome profile from the dinoflagellate *Lingulodinium polyedrum* has allowed an in depth analysis of histone and histone modifying genes in a single species. We report here that this species expresses a full set of core histone genes as well as a variety of histone modifying enzymes and histone chaperone proteins at the RNA level. Despite the fact we have not been able to detect histone proteins in *Lingulodinium* extracts the presence and highly conserved sequence of these genes indicates that, in contrast to what has been previously thought, dinoflagellates do indeed have histones.

2.3. Materials and Methods

2.3.1. Cell Culture

Lingulodinium polyedrum cultures (formerly *Gonyaulax polyedra*; strain CCMP1936) were obtained from the Provasoli-Guillard Culture Center for Marine Phytoplankton (Boothbay Harbor, Maine) and grown in a modified seawater medium (f/2) [399] at constant temperature ($19 \pm 1^\circ\text{C}$) in 12-h light/12-h dark cycles using cool white fluorescent light at an intensity of $50 \mu\text{mol photons m}^{-2}\cdot\text{s}^{-1}$. The beginning of light period is defined as LD 0, and the beginning of the dark period as LD 12. Cultures were grown to a cell density of 12-14,000 cells/mL. The samples were collected from the middle of the dark phase (LD 18) by filtering on Whatman 541 paper supported by a Buchner funnel, and either used immediately or frozen in liquid nitrogen and stored at -80°C until further use.

2.3.2. Acid Extraction of proteins

Histone proteins were obtained by trichloroacetic acid (TCA) precipitation of the acid soluble protein fraction as described previously [400, 401] with some minor modifications. After washing with 10 volumes of fresh f/2 medium the cells were suspended in ice-cold acid extraction buffer containing 10 mM HEPES pH 7.9, 1.5 mM MgCl_2 , 10 mM KCl, 0.5 mM DTT and 1.5 mM Phenyl methyl sulfonyl fluoride supplemented with 1 X EDTA-free Protease Inhibitor (from Roche) and HCl at a final concentration of 0.25 M. Cells were broken by three one-minute treatments in a bead beater with Zirconium beads at 4°C . The lysate was then incubated on a rotator for 1 hour at 4°C . Insoluble cell debris was removed by two sequential centrifugations at $11,000 \times g$ for 10 and 5 minutes, each at 4°C , and the supernatant retained. To this acid soluble fraction, 100 % TCA was added drop by drop with simultaneous mixing by inverting the tubes several times until a final concentration of 33% (v/v) TCA was reached. The solution was then incubated overnight at 4°C and the acid soluble proteins were obtained by centrifugation at $16,000 \times g$ for 10 minutes at 4°C . To remove the acid, the pellet was carefully washed three times with ice-cold acetone using

centrifugation at 16,000 X g for 5 minutes at 4 °C after each wash. The final pellet was air dried and dissolved in appropriate amount of ddH₂O.

As a positive control, *Saccharomyces cerevisiae* (budding yeast) was cultured in 100 ml of 2X YPAD medium at 30 °C to mid-log phase ($A_{600} = 0.6$). Cells were then harvested by centrifugation at 4 °C for 5 min at 2,000 X g and washed once with 10 volumes of ice-cold sterile Phosphate buffered saline (pH 7.3). All the procedures after this were the same as described above for *Lingulodinium* cells. All protein concentrations were measured using the Bradford assay (Bio-Rad).

2.3.3. SDS-PAGE and Immunoblotting

Lingulodinium and *Saccharomyces* acid soluble proteins along with molecular weight markers (Low Range-BIORAD) were resolved by SDS-15% Polyacrylamide gel electrophoresis (PAGE) as previously described [401]. To compare the protein profiles after electrophoresis, some gels were stained with Coomassie Blue, while others were used for western blotting. Western blotting was performed using commercial rabbit polyclonal antibodies for histones H3 (ab 1791, Abcam, USA) and H2B (sc-10808, Santa Cruz Biotechnology, USA). For Immunoblotting, the proteins from gels were transferred to the Hybond-P PVDF membranes (Amersham Biosciences) using the Transblot SD Semi-Dry Electrophoretic transfer cell (Bio-Rad) following the manufacturer's protocol. After blocking the membranes with 5% Non-fat dry milk in Tris-buffered saline buffer supplemented with 0.05% Tween-20, immuno-reaction was performed with H3 (1:5000) and H2B (1:1000) antibodies in the same buffer. After secondary antibody reaction and subsequent washings, the blots were developed with Chemiluminescent substrate (Millipore) and were exposed to the ImageQuant LAS 4000 (GE Healthcare) to capture the chemiluminescence.

In order to test the commercial H3 antibody for cross-reaction with the *Lingulodinium* protein, a tagged version of our H3 was expressed in bacteria. The H3 sequence was cloned by PCR using primers based on the transcriptome sequence (forward primer 5'-CATTACGCCTGACGCTGTCTACGTGC-3' and reverse primer 5'-

GTTAGCGTCTGCTGCTGACGGCTTC-3') from a 1st strand cDNA sample prepared from Trizol (Invitrogen) extracted RNA using a reverse transcription reaction catalyzed by MMLV RT (Clontech) and the 5' CDS primer A of the SMARTer RACE cDNA Amplification kit (Clontech). A second PCR, performed on the first PCR product using the forward primer 5'-TCAGTCg gatccATGGCCCGCACGAAGCAG-3' (containing a BamH1 site indicated by small letters) was used to allow directional cloning into the BamH1 and Sma1 restriction sites of the bacterial expression vector pQE30 (Qiagen). The cloned H3 was sequenced to confirm the correct reading frame and used to transform electrocompetent XL1 blue host cells. A single colony grown on LB-agar containing tetracycline and ampicillin was inoculated into 5 mL of the same medium and left to grow overnight at 37 °C. One mL of the overnight culture was used to inoculate twenty ml of fresh prewarmed (37 °C) LB medium with antibiotics and grown with vigorous shaking at 37 °C until OD⁶⁰⁰ of 0.5. H3 expression was induced by adding IPTG to a final concentration of 1 mM and the culture was grown for another 4 hours with shaking at 37 °C. One ml of this culture was centrifuged at 5000 x g for 3 minutes at 4 °C and the cell pellet resuspended directly in 50 µl SDS-PAGE sample buffer and heated at 95 °C for 5 min. The samples were centrifuged to remove debris and 30 µl of sample was loaded onto a 15% polyacrylamide gel. XL1 blue cells containing an empty vector were used as a control. Electrophoresis, transfer and immunoblotting were carried out as above.

2.3.4. Mass Spectrometric analysis

The total acid soluble protein pellet in acetone was also used for mass spectrometric analysis. Also, after fractionating the yeast and *Lingulodinium* acid soluble proteins in SDS 15% PAGE, the gels were stained with Coomassie Blue and several regions were excised from the gel, both from the *Lingulodinium* and yeast samples. The excised bands were destained and sent to the proteomic facility of l'Institut de recherche en immunologie et en cancérologie (IRIC) in Montreal, Canada. The tryptic digestion and LC-MS/MS sequencing for both the total acid extracted proteins and fractionated gel-excised bands were performed at the IRIC.

2.3.5. Bioinformatic and Phylogenetic Analysis

The sequences for histones and histone modifying enzymes reported here were retrieved from a *Lingulodinium* transcriptome assembled from roughly 300 million 76 bp Illumina paired end reads combined from several times under a LD cycle and conditions (manuscript in preparation, GenBank Accession numbers JO692619 through JO767447). The Illumina sequencing and assembly was performed at the Genome Quebec sequencing facility. The number of reads corresponding to each histone sequence was determined for RNA samples prepared over LD 6 and LD 18 cell cultures and reported as number of histone reads present per million. The number of reads for the histone sequences in the wild potato *Solanum chacoense* was retrieved from a similar project undertaken concurrently with the *Lingulodinium* samples.

Phylogenetic analysis was performed using an online tool obtained from the website www.phylogeny.fr [402]. Our workflow used the software MUSCLE to align the histone sequences, curation by GBlocks, phyML bootstrapping (100 times) to construct the tree and TreeDyn to visualize the tree. The same workflow was followed for all the phylogenetic analysis.

2.4. Results

2.4.1. All core histone and many histone modifying enzyme sequences are present in the *Lingulodinium* transcriptome

Analysis of a recent Illumina sequencing run (manuscript in preparation) identified the entire set of core histones, namely H2A, H2B, H3 and H4 from the dinoflagellate *Lingulodinium polyedrum* (Table 2.1.). Partial splice leader sequence [266] was recovered from at least one of each histone sequences and all the sequences are GC-rich, a common characteristic of many dinoflagellate sequences [398]. In addition, *L. polyedrum* also expresses genes encoding enzymes that post-translationally modify histones, such as histone lysine methyltransferase (KMT), histone arginine methyltransferase (PRMT), histone acetyltransferase (KAT) and also histone deacetylases (from both HDAC and sirtuin 2 superfamilies) (Table 2.2.0). We also found histone chaperone proteins (NAP, ASF1-like), which assist in nucleosome formation and chromatin remodelling (Table 2.2.). *Lingulodinium* thus expresses a wide range of genes responsible for making and modifying nucleosomes.

2.4.2. Phylogenetic grouping identifies at least two major variants of all histone sequences within *Lingulodinium*

The *Lingulodinium* transcriptome contains at least two variants of each histone sequence. We thus performed phylogenetic analyses to provide insight into the relationship between the different histone variants. Among the three H2A sequences retrieved, two belong to class H2A.X while the other groups with the eukaryotic H2A.Z proteins (Figure 2.1.). This is the first report of a Z – like variant of histone H2A in any dinoflagellate. The two H2A.X sequences, JO760634 and JO759158, both contain a signature SQEF motif at the C-terminal end that is common to all dinoflagellate H2A.X sequences known so far [274] and as expected, all the dinoflagellate H2A.X variants cluster together. Interestingly, the two *L. polyedrum* H2B proteins belong to two different clades, one common to other dinoflagellate H2B (JO720817) and the other (JO694219) grouping within the superphylum Alveolata along with the ciliates and apicomplexans (Supplementary Figure 2.S1.).

Similarly, there are two well supported clades of H3 sequences, one phylogenetically indistinguishable from other eukaryotic H3 sequences and the other divergent (JO753891) form also found in *Pyrocystis lunula* (Supplementary Figure 2.S2.). Unfortunately, there is insufficient phylogenetic resolution to determine the origin of the *Lingulodinium* H4 proteins (supplementary Figure 2.S3.). In general, however, it seems *Lingulodinium* contains not only a dinoflagellate specific histone but also an additional sequence with a more divergent origin.

2.4.3. Histone mRNAs abundance levels are uniform throughout

Replication – dependent histone sequences tend to accumulate during the S-phase of the cell cycle. In *Lingulodinium*, S-phase begins in the middle of the dark phase (LD18) for cells grown under a 12:12 L:D cycle [240]. We therefore compared the number of sequence reads in a sample from mid-day (LD6) with the LD18 sample. No significant variation in the mRNA abundance between the day and night is supported by the data (Table 2.1.). In general, all the histone mRNAs seem to be of low abundance. By way of comparison, we found *Lingulodinium* histone mRNA abundance to be roughly 5 to 25-fold lower than in the plant *Solanum chacoense*.

2.4.4. Histone protein accumulation is below current detection limits

To reconcile the apparent lack of nucleosomes in dinoflagellates with the expression of all core histone transcripts in *Lingulodinium*, we evaluated the extent of histone protein accumulation using more sensitive techniques than those used previously. As shown previously [185, 188] acid extracted proteins from *Lingulodinium* do not have the typical pattern of histones such as found in yeast extracts using SDS PAGE followed by Coomassie blue staining (Figure 2.2.). We used LC-MS/MS to analyze the *Lingulodinium* acid extracted proteins and both the entire acid extracted protein fraction as well as the acid extracted proteins that had been further fractionated by SDS-PAGE into the size range of yeast histones was tested. None of the histone core sequences from *Lingulodinium* were found in any of our samples although we were able to detect *Lingulodinium* histone-like protein, as expected (Table 2.3., Supplementary Table 2.S1.). As a control, the same experiment was performed with an acid extracted fraction of a yeast extract, and histone sequences H2A and H2B were readily detected (Table 2.3.).

In a separate approach, we tested *Lingulodinium* acid extracted proteins for a cross reaction with histone antibodies. We first tested a commercial anti-H3 directed against an epitope that shared 92% sequence identity with the *Lingulodinium* sequence. This antibody detected the yeast H3 with as little as 0.07 µg of total acid extracted protein, whereas as much as 20 µg of acid extracted protein from *Lingulodinium* did not show a reaction with any protein corresponding in size to the yeast H3 band (Figure 2.3.). The high protein load of *Lingulodinium polyedrum* extracts show cross-reacting proteins with a significantly different mobility from the yeast H3, but the identity of these proteins is unknown. We also tested an antibody raised against the full length H2B sequence of mammalian origin, and again the antibody was unable to detect any band corresponding in size to that of yeast H2B (Supplementary Figure 2.S4.). Again, at high concentrations of protein the antibody showed a cross reaction with a band with reduced mobility (~ 30 kD) whose identity is also unknown. As a caveat, however, the H2B used to generate this antibody is only 63% similar to the predicted protein produced by the *Lingulodinium polyedrum* H2B.

2.5. Discussion

Nucleosomes are the basic structural and functional unit of chromatin in most eukaryotes, and are formed when roughly 150 bp of DNA wrap around a histone octamer (two each of H2A, H2B, H3 and H4). Dinoflagellates differ from other eukaryotes in that DNA spreads do not show nucleosomes [184, 403, 404], 150 bp DNA fragments of DNA are not protected from micrococcal nuclease digestion [187, 405] and gels of basic proteins do not show the typical histone protein pattern [188]. This general rule for dinoflagellates has only two known exceptions, the binucleate dinoflagellates such as *Peridinium balticum* (which have both typical eukaryotic and dinoflagellate nuclei) [406, 407] and members of the endoparasitic *Perkinsus* whose nuclei resemble those in a typical eukaryotic cell [197]. *Perkinsus marinus* is considered to be the ancestor of the dinoflagellate lineage [386], and not only contains all the core histone sequences [408] but lacks the HLPs found in other dinoflagellates.

Recently, high throughput sequencing has revealed that environmental samples of dinoflagellates transcripts contain not only the four core histones, but also two histone modification proteins and a NAP [274]. However, while these sequences are clearly dinoflagellate in origin, based on the distinguishing SL sequence at the 5' end [266], it is not clear if they are all expressed in the same species. We show here that a single species of dinoflagellate expresses all the core histone (Supplemental Figures 2.S5. – 2.S8.) as well as a wide range of histone modifying enzymes and histone chaperone proteins (Table 2.1. and 2.2.). Furthermore, the gene profile is surprisingly complex, with at least two different variants of predicted histone sequence, one relatively close to other eukaryotic histones and the other more divergent (Figure 2.1. and Supplementary Figures 2.S1. – 2.S3.).

Among the core histones, histone H2A has several subtypes including H2A.1 and 2, H2A.X and H2A.Z. These subtypes each contain signature sequence elements that have been conserved throughout evolution and allow them to be readily identified [409, 410]. In mammals, all the major variants of H2A are present in varying proportions, whereas lower

eukaryotes often replace the more common H2A.1 and 2 subfamily with H2A.X [411]. *Lingulodinium* also contains the H2A.X variant and in addition, an H2A.Z-like subtype previously unreported in dinoflagellates (Figure 2.1.). These subtypes are thought to have specific functions, with H2A.X directly involved in DNA repair and genome integrity, which requires the phosphorylation of the C-terminal serine (S) of the SQ(D/E)(M/Y/F) motif [412], and H2A.Z involved in chromosome segregation, cell cycle progression and regulation of expression of cyclin genes, which is mediated by the H2A.Z localized in the promoter regions of these genes [413]. For the H2B and H3 histones, *Lingulodinium* maintains a general eukaryote form in addition to a divergent form common to other dinoflagellates (Supplementary Figures 2.S1. and 2.S2.). Interestingly, two of the three H3 sequences in *Lingulodinium* conserve the key post-translational modification sites K4, K9, K27, K36 and K79 [414], while the other divergent forms lack the K27/K36, as in *Pyrocystis* H3 and K79 as in *Karlodinium* H3. For H4, we found two sequences (Supplementary Figure 2.S3.), all with a conserved K20 site, which has been linked to transcription repression upon methylation [414]. Thus, the presence of all core histones, the conservation of sites typically modified, as well as the presence of histone modifying enzymes in the transcriptome (Table 2.2.), all suggest that *Lingulodinium* should accumulate histone proteins.

We had originally anticipated that the amount of histone proteins expected for *Lingulodinium* could be estimated by assuming that the amount of protein produced from a transcript will be proportional to the amount of message independent from the organism in which the transcript is found. We therefore compared the amount of histone transcripts in *Lingulodinium* with those of the plant *Solanum chacoense*, as RNA samples from both were prepared, sequenced and analysed concurrently. In general, the abundance of histone messages in *Lingulodinium* is roughly 30 times less than that in *S. chacoense* (Table 2.1.) and roughly 60-fold less than that reported for yeast [415]. However, immunoblotting was unable to detect H3 in *Lingulodinium*, even when the amount of *Lingulodinium* protein was 300 times greater than yeast. Furthermore, histone proteins were not detected by mass spectrometry (Table 2.3.), either in total or gel fractionated acid soluble extracts, even though other proteins detected in the extracts had similar transcript levels as the *Lingulodinium* histones (Table 2.4.). Thus, it seems histone abundance may be lower than would be

predicted. It might also be of interest to test different extraction procedures for histones to see if this aids detection.

Histone modification has been linked to several functions such as chromatin remodelling and epigenetic regulation [416], and thus the finding that the *Lingulodinium* transcriptome also contains histone acetyltransferase and deacetylase enzymes as well as methyltransferases (Table 2.2.) supports a role for histones in regulating gene expression. However, it must be noted that while histone deacetylases have a strong link to gene repression and heterochromatin formation [396, 417, 418], they can also target non-histone proteins and regulate DNA binding affinity, protein stability and protein-protein interaction, as well as modulate enzyme activity [419]. Sirtuin family proteins, deacetylases overrepresented in our transcriptome, were also reported in prokaryotes and archaea [420] where they function to regulate metabolism through important enzymes like acetyl-CoA synthetase [421]. Similarly, the SET domain K-methyltransferase that methylates histones can also methylate diverse proteins such as cytochrome *c* and the large subunit of Rubisco [422, 423]. A SET domain histone methyltransferase (NUE) has been reported in the pathogenic bacteria *Chlamydia trachomatis* [424]. Thus, it is possible the histone modifying enzymes in *Lingulodinium* might modify proteins other than the core histones. One prospective substrate could be the *Lingulodinium* HLPs, which have been reported to be acetylated [115]. Similarly, histone chaperone proteins also have important alternative roles other than those related to nucleosome assembly. NAP family proteins specifically interact with B-type cyclin [425, 426] and play a role in regulating cell cycle [427]. It would be of interest to determine if any of the histone modifying enzymes are, unlike the histones themselves, detectable immunologically.

The abundance of histone mRNA in *Lingulodinium* is between 5- and 25-fold lower than in the higher plant *Solanum chacoense* depending on the histone (Table 2.1.). In eukaryotes, histones are found in both replication-dependent and replication-independent classes [428], with the mRNA abundance of replication-dependent histones coupled to the cell cycle as expected [429]. Transcriptional and posttranscriptional regulation can result in a 15- to 30-fold increase in mRNA accumulation with a peak during mid S phase [430, 431]. A

comparison of histone mRNA levels at LD 6 and LD 18 (Table 2.1.) does not show preferential abundance during the LD 18, the peak of S-phase in *Lingulodinium* [240, 241]. Thus, histone transcript accumulation is independent from the cell cycle in *Lingulodinium*.

Our results with *Lingulodinium* show that all core histone transcripts are present in a single species. Although histone protein levels remain below our current limit of detection, the presence of all four core histone transcripts, the conservation of their sequence, and the presence of a large number of histone modifying enzymes all support the hypothesis that dinoflagellates have histones.

Table 2.1. Description of histone sequences and their relative abundance in *Lingulodinium*

The preliminary identification of histone sequences was carried out using the blastX function incorporated within the Blast2GO software using a cut-off value of e^{-10} . Once identified, all prospective sequences were thoroughly analysed.

| Histone | <i>S. chacoense</i> | <i>L. polyedra</i> | GC | | |
|---------|---------------------|--------------------|---------|-----------------|-----------------|
| | | Sequence ID | content | LD 6 | LD 18 |
| | (reads/million) | | | (reads/million) | (reads/million) |
| H2A | 67 | JO760634 | 64% | 4 | 4 |
| | | JO759158 | 69% | 2 | 2 |
| | | JO731189 | 55% | 6 | 6 |
| H2B | 30 | JO694219 | 65% | 2 | 2 |
| | | JO720817 | 68% | 1 | 2 |
| H3 | 124 | JO722862 | 66% | 2 | 3 |
| | | JO740554 | 75% | 1 | 1 |
| | | JO753891 | 65% | 2 | 2 |
| H4 | 63 | JO717937 | 70% | 2 | 2 |
| | | JO719134 | 66% | 3 | 3 |

Table 2.2. Description of histone modifying enzymes and histone chaperones based on blastX alignments.

| Protein ID | Hit protein family | Hit Accession Number | E-Value | Similarity | GC content |
|------------|--------------------|----------------------|--------------------|------------|------------|
| JO734372 | KAT, ELP3 | XP_002773536.1 | 1 e ⁻⁷¹ | 67% | 67.9% |
| JO732038 | KAT, ELP3 | XP_002773536.1 | 6 e ⁻⁷² | 78% | 65.9% |
| JO710977 | HDAC | XP_001758783.1 | 3 e ⁻⁷⁰ | 60% | 66.9% |
| JO734243 | HDAC | BAB10370.1 | 9 e ⁻⁴⁵ | 67% | 66.7% |
| JO742233 | HDAC | XP_001625421.1 | 1 e ⁻⁷¹ | 68% | 72.3% |
| JO743978 | HDAC | XP_002514660.1 | 1 e ⁻⁷¹ | 67% | 68.5% |
| JO724091 | HDAC, SIR2 | XP_003057268.1 | 2 e ⁻⁸² | 67% | 67.2% |
| JO726045 | HDAC, SIR2 | XP_002508530.1 | 1 e ⁻⁷⁶ | 70% | 73% |
| JO733933 | HDAC, SIR2 | XP_003057268.1 | 4 e ⁻⁷⁵ | 67% | 69.3% |
| JO726372 | KMT, SET | XP_003195141.1 | 2 e ⁻³⁰ | 51% | 68.4% |
| JO694016 | KMT, SET | XP_002785418.1 | 4 e ⁻¹⁷ | 49% | 73.5% |
| JO752203 | PRMT | NP_001150868.1 | 5 e ⁻⁶⁴ | 56% | 69.3% |
| JO723144 | PRMT | NP_001003645.1 | 6 e ⁻⁴⁹ | 60% | 65.6% |
| JO735881 | PRMT | XP_001945590.2 | 8 e ⁻⁶² | 62% | 69.4% |
| JO747341 | NAP | XP_002764795.1 | 2 e ⁻³² | 55% | 64.3% |
| JO745850 | NAP | XP_002764795.1 | 6 e ⁻³⁴ | 50% | 70.1% |
| JO738268 | NAP | XP_002764795.1 | 2 e ⁻²⁶ | 54% | 61.7% |
| JO761496 | NAP | XP_002764795.1 | 2 e ⁻³⁸ | 57% | 65.7% |
| JO748499 | ASF1-like | XP_758562.1 | 1 e ⁻¹⁹ | 57% | 69.3% |
| JO750428 | NAP-C | ADE76527.1 | 6 e ⁻⁶³ | 49% | 69.3% |

Table2.3. Proteins found by LC-MS/MS sequencing of total acid soluble proteins from *Lingulodinium* and yeast

| | Type of Protein | No. Proteins (≥ 2 peptides) | Confidence | Species Hit |
|---------------------|--|--------------------------------------|--|--|
| <i>L. polyedrum</i> | Histone like protein | 1 | 9 e^{-26} | <i>L. polyedrum</i> |
| | Perilipin-4 | 3 | 7 e^{-23} | <i>Bos taurus</i> |
| | Photosystem II 12 kDa extrinsic protein | 3 | 3 e^{-37} ; 4 e^{-36} ; 2 e^{-35} | <i>Heterocapsa triquetra</i> <i>Leishmania mexicana</i> |
| | Kinesin-K39 | 1 | 5 e^{-13} | <i>mexicana</i> |
| | Elongation factor-1 α | 2 | 0 | <i>H. triquetra</i> |
| | Malate dehydrogenase | 1 | 1 e^{-115} | <i>H. triquetra</i> |
| | Peptidoglycan domain containing protein | 1 | 5 e^{-09} | <i>Tetrahymena thermophila</i> |
| | H2A-1 | 1 | 0 | <i>S. cerevisiae</i> |
| | H2B-1 | 1 | 0 | <i>S. cerevisiae</i> |

Table 2.4. mRNA abundance of expressed proteins detected by LC-MS/MS in an acid-extracted protein fraction

| Accession number | LD6 reads | LD 18 reads | Blast hit (E-value) |
|------------------|-----------|-------------|---|
| JO757244 | 1 | 1 | Unknown |
| JO711184 | 3 | 3 | Unknown |
| JO741176 | 1 | 1 | 30S ribosomal protein S11 (4e-05) |
| JO735533 | 2 | 2 | Unknown |
| JO698965 | 6 | 5 | Unknown |
| JO760395 | 4 | 3 | U1 small nuclear ribonucleoprotein A-like (3e-63) |
| JO764129 | 6 | 5 | Unknown |

Figure 2.1. Two variants of Histone H2A in *Lingulodinium*

The cladogram of histone 2A.X and Z variants shows representatives from mammals, plants, fungus and members of the superphylum Alveolata. The representative sequences were obtained from Pubmed database and bear the first three letters from genus followed by two letters from species. The values in red at each node indicate the respective Bootstrap support value. *Lingulodinium* sequences are coloured in blue. The representative classes are *Homo sapiens* (Hom_sa), *Rattus norvegicus* (Rat_no), *Danio rerio* (Dan_re), *Xenopus tropicalis* (Xen_tr), *Arabidopsis thaliana* (Ara_th), *Zea mays* (Zea_ma), *Cicer arietinum* (Cir_ar), *Saccharomyces cerevisiae* (Sac_ce), *Candida dubliniensis* (Can_du), *Toxoplasma gondii* (Tox_go), *Neospora caninum* (Neo_ca), *Perkinsus marinus* (Per_ma), *Alexandrium tamarense* (Ale_ta), *Cryptocodinium cohnii* (Cry_co).

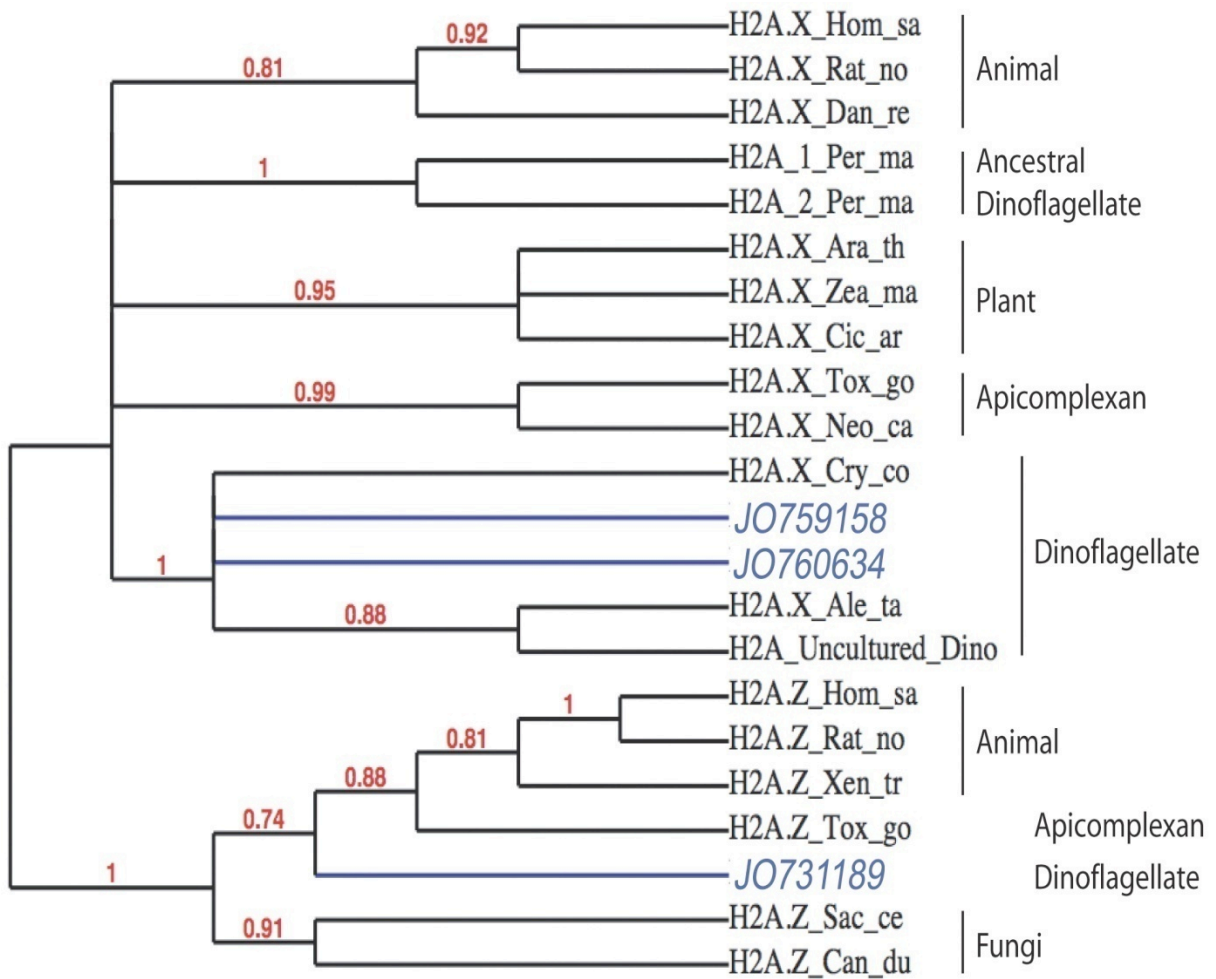


Figure 2.2. The acid soluble protein profiles of *Lingulodinium* and Yeast differ

A Coomassie blue stained gel containing roughly equivalent amount of acid extracted proteins from *Lingulodinium* and yeast in SDS-15% PAGE is shown here. The regions of the gel corresponding to yeast histones were excised and analysed by LC-MS/MS.

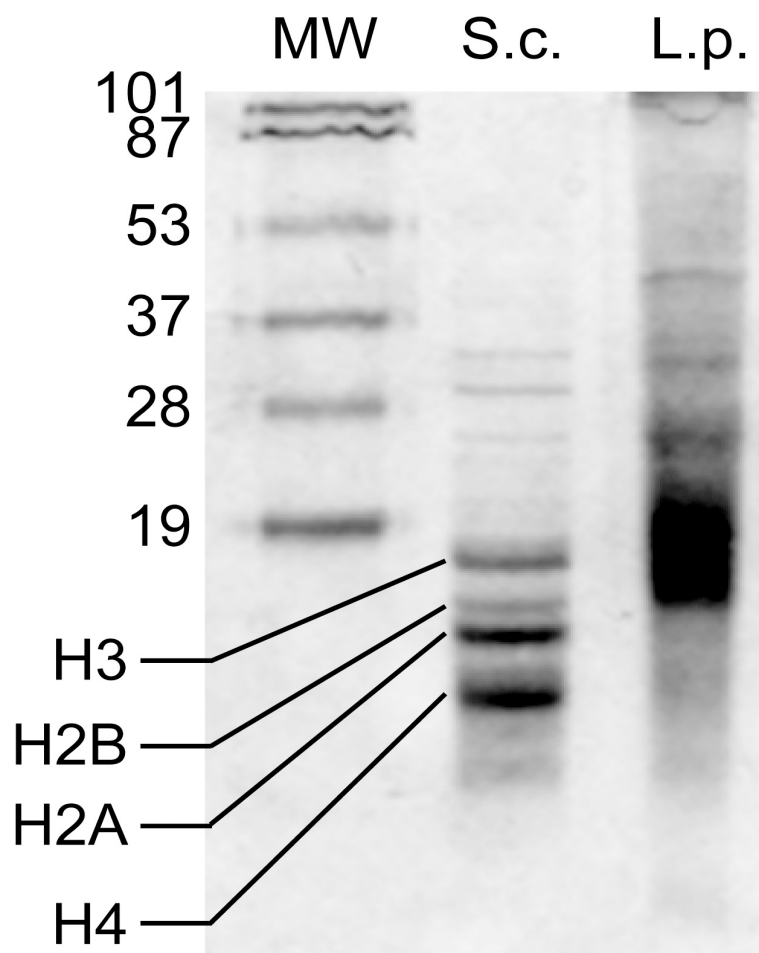
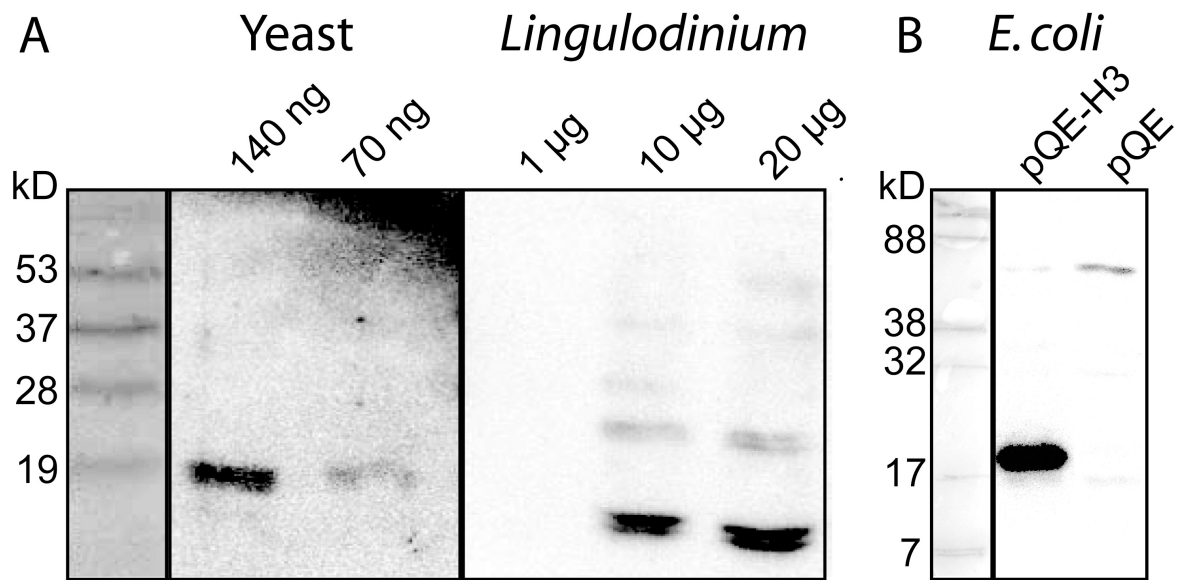


Figure 2.3. Histone H3 protein levels in *Lingulodinium* are below current immunodetection limits

(A) Acid extracted proteins electrophoresed on SDS-15% PAGE were subjected to Western blot analysis using a commercial H3 antibody. For the yeast and *Lingulodinium* samples, the value above each lane indicates the amount of protein loaded in micrograms, and the samples were run and treated with antibodies concurrently. No signal is detected in the *Lingulodinium* sample at a position corresponding to the yeast H3. (B) Western blots, performed using the same anti-H3 and an H3-expressing *E. coli* strain or an *E. coli* strain containing only the empty vector, demonstrate cross-reaction of the antibody with the *Lingulodinium* H3.



Supplementary Figure legends

Figure 2.S1.Cladogram of histone H2B

The cladogram of histone sequences shows representatives from mammals, plants, fungus and members of the superphylum Alveolata. The representative sequences were obtained from Pubmed database and bear the first three letters from genus followed by two letters from species. The values in red at each node indicate the respective Bootstrap support value. *Lingulodinium* sequences are coloured in blue. The representative sequences other than those present in Figure 2.1 are *Ricinus communis* (Ric_co), *Physcomitrella patens* (Phy_pa), *Neurospora crassa* (Neu_Cr), *Amphidinium carterae* (Amp_ca), *Karlodinium veneficum* (Kar_ve), *Thalassiosira pseudonana* (Tha_ps), *Phaeodactylum tricornutum* (Pha_tr), *Paramecium caudatum* (Par_ca), *Tetrahymena thermophila* (Tet_th), *Tetrahymena pyriformis* (Tet_py), *Babesia bovis* (Bab_bo), *Plasmodium falciparum* (Pla_fa), *Plasmodium chabaudi* (Pla_ch), *Cryptosporidium muris* (Cry_mu).

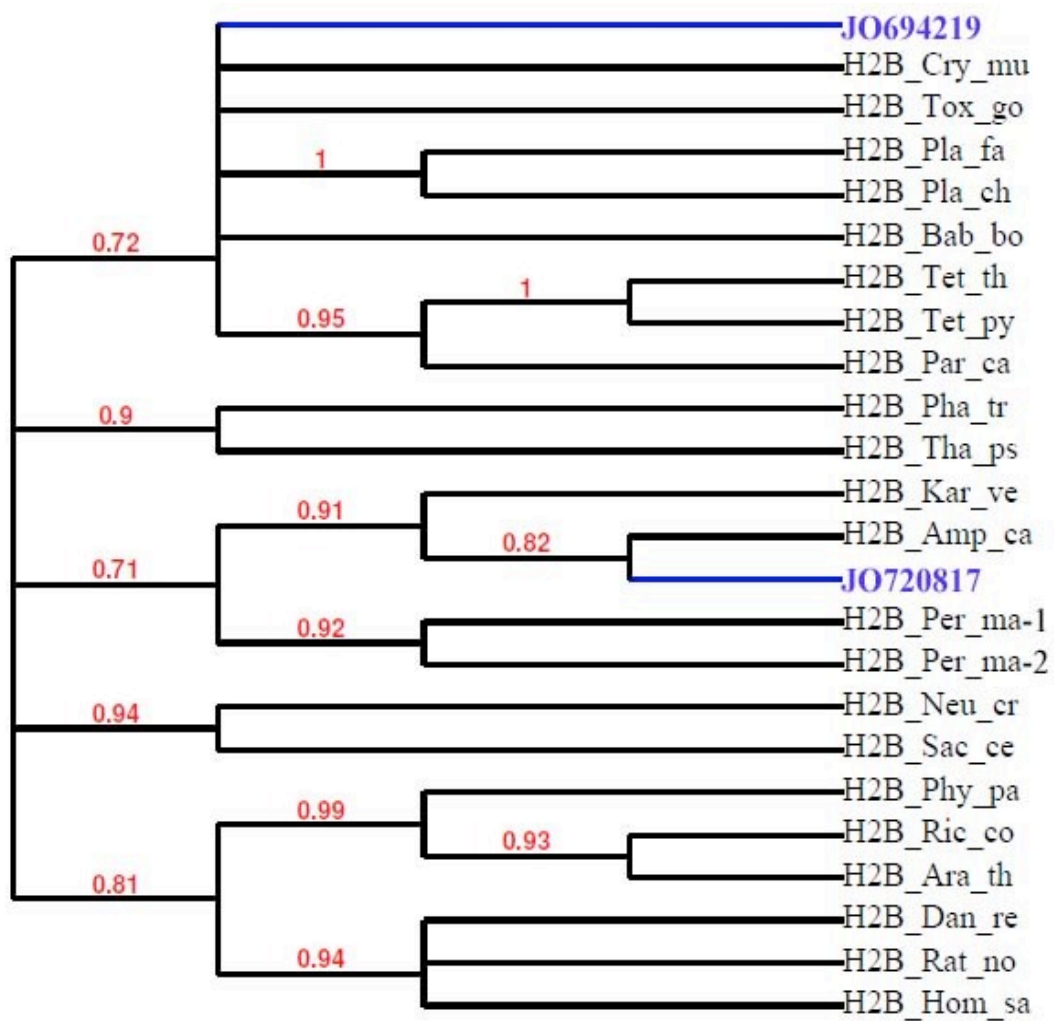


Figure 2.S2. Cladogram of histone H3

The cladogram of histone sequences shows representatives from mammals, plants, fungus and members of the superphylum Alveolata. The representative sequences were obtained from Pubmed database and bear the first three letters from genus followed by two letters from species. The values in red at each node indicate the respective Bootstrap support value. *Lingulodinium* sequences are coloured in blue. The representative sequences other than those present in Figure 2.1 and 2.S1 are *Drosophila melanogaster* (Dro_me), *Chlamydomonas reinhardtii* (Chl_re), *Cryptosporidium parvum* (Cry_pa), *Plasmodium vivax* (pla_vi), *Pyrocystis lunula* (Pyr_lu), *Karlodinium micrum* (Kar_mi).

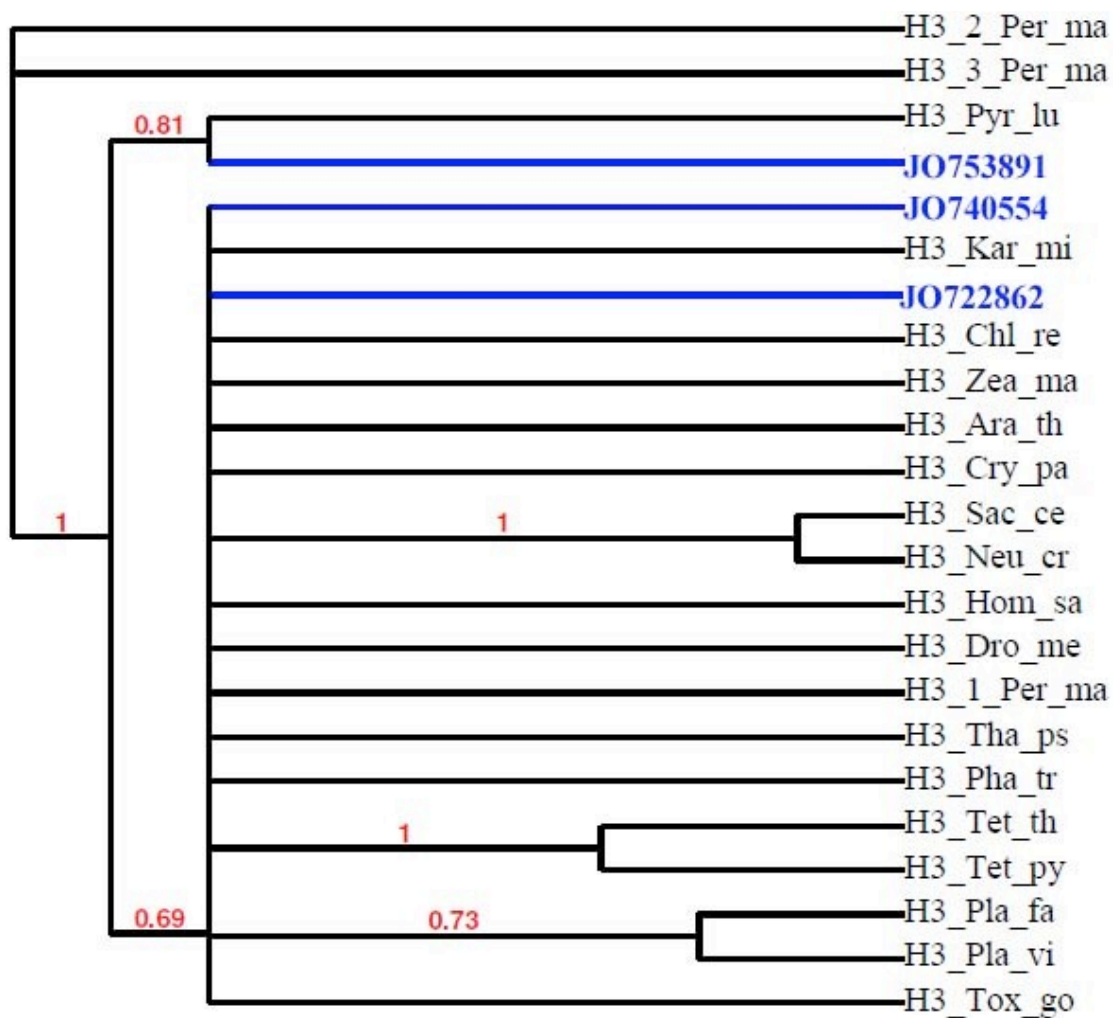


Figure 2.S3. Cladogram of histone H4

The cladogram of histone sequences shows representatives from mammals, plants, fungus and members of the superphylum Alveolata. The representative sequences were obtained from Pubmed database and bear the first three letters from genus followed by two letters from species. The values in red at each node indicate the respective Bootstrap support value. *Lingulodinium* sequences are coloured in blue. The representative sequences other than those present in Figure 2.1, 2.S1 and 2.S2 are *Paramecium tetraurelia* (Par_te), *Plasmodium yoelii* (Pla_yo).

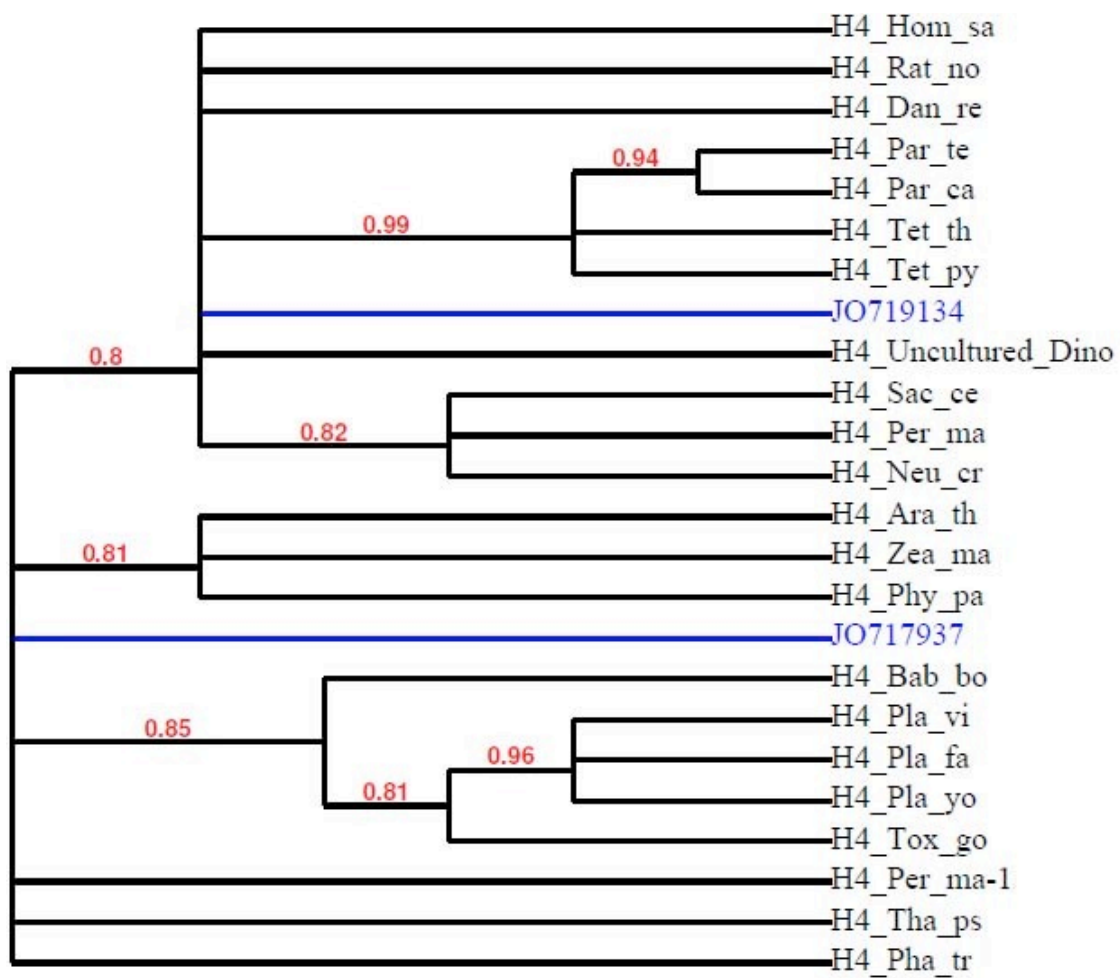


Figure 2.S4. Histone H2B protein is not detected in *Lingulodinium*

Western blotting with H2B antibody is shown here. The amount of protein (in micrograms) per lane is written above each lane.

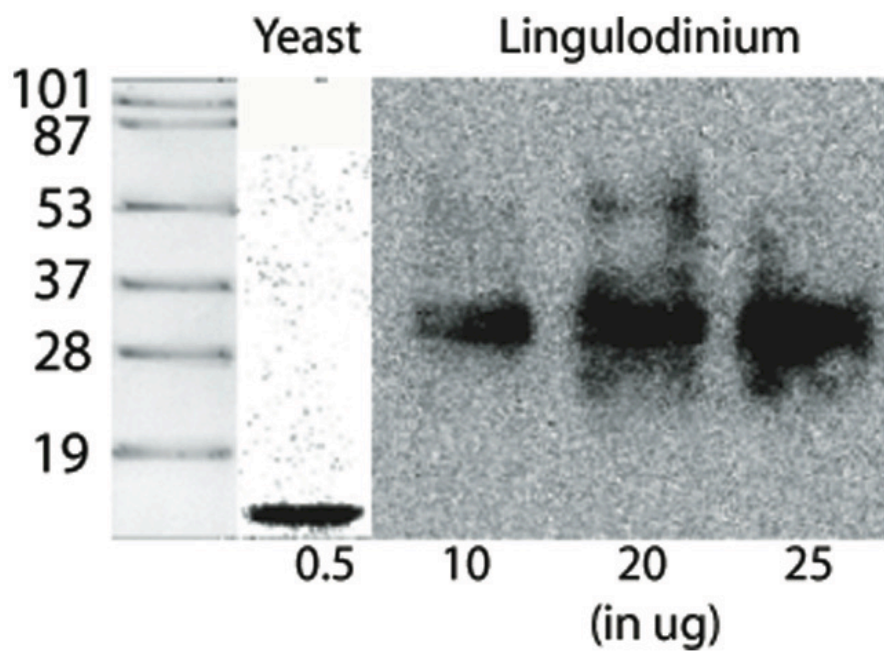


Figure 2.S5. Alignment of H2AX sequences.

Multiple sequence alignment of histone H2A from human, *Alexandrium* and *Lingulodinium* is shown

| | | | | | | |
|--------------|---|----------|--------|---------|---------|-----------------------------------|
| | 1 | 10 | 20 | 30 | 40 | 50 |
| H2A.X_Ale_ta | | | MITNLI | RLTIGNL | ALAEAK | NSARGSAAGKQKGSS---- |
| JO753891 | | MDDDF | EDKGT | KKIVG | AKTPIG | SGLKGSAGVGDFDKGSKKVVM |
| H2A.X_Hom_sa | | | | | MSGR | GKTGGKAR----- |
| JO759158 | | MQA | SQEAP | QEEFE | DEQQDA | AEVDEEESQDDEEEDEGAAEGSEKPRPVIKD |
| JO760634 | | MPKA | HEQE | EEEE | EQDAE | VEDEEEDGEGEDAEEEEAGGAATSRRKRPVAKD |
| | | 60 | 70 | 80 | 90 | 100 |
| H2A.X_Ale_ta | | ----- | SKK | AGKPV | QGEK | KKHQ----- |
| JO753891 | | AAG | TKAP | IMAEG | LTIKDK | AVKEKDKASKGKDQDKDKLKKGKKNPKWGKTD |
| H2A.X_Hom_sa | | ----- | ----- | ----- | AKAK | ----- |
| JO759158 | | SAG | ----- | V | KKGSS | STNPERQKS----- |
| JO760634 | | PAG | ----- | VKKK | QAPREG | ERTKA----- |
| | | 110 | 120 | 130 | 140 | 150 |
| H2A.X_Ale_ta | | --- | SLSS | KAGLV | FPVQR | FAKALRKG-GYCKRLAAGGSIYLTAVIEYITA |
| JO753891 | | GKAL | SR | SARAGV | QFPVGRV | HRFLKQYITHNMRVGGTAGVYVAVMEYLSA |
| H2A.X_Hom_sa | | --- | SRSS | RAGLO | FPVGRV | RLLRKG-HYAERVGAGAPVYLA |
| JO759158 | | --- | SAN | ARAGLV | FPVHR | FAKALRKG-GYAKRLAMGGSVYLTAVIEYITA |
| JO760634 | | --- | APN | KRAGLV | FPVHR | IAKALRKG-GYAKRLAVGGSVYMTAVIEYLTA |
| | | 160 | 170 | 180 | 190 | 200 |
| H2A.X_Ale_ta | | EILELAGN | SAKEA | RKQRI | IPRHI | QLAIRNDEELNKYLSRVTTITGGGVLPN |
| JO753891 | | EVLELAGN | VAKDL | KT | KRIT | PRHLQLAIRGDEELDALTIKATTIAGGGVVPH |
| H2A.X_Hom_sa | | EILELAGN | AARDN | KK | TRIT | PRHLQLAIRNDEELNKLLGGVTIAQGGVLPN |
| JO759158 | | EILELAGN | AAKDQ | KKQRI | IPRHI | QLAIRNDEELNKYMSNVTISGGGVLPN |
| JO760634 | | EILELAGN | AAKDM | KKQRI | VPRHI | QLAIRNDEELNKYMSNVTISGGGVLPN |
| | | 210 | 220 | 230 | 232 | |
| H2A.X_Ale_ta | | IHTV | LLPKV | ----- | MPA | KEGASGSFSQEF |
| JO753891 | | IDK | SLLTKK | GKKKAT | --DA | SAMVMSASVGGEF |
| H2A.X_Hom_sa | | IQAV | LLPKKT | SATV | GPKA | PSGGKKATQASQEF |
| JO759158 | | IHSV | LLPKKA | ----- | V | PGKDGVSASFQEF |
| JO760634 | | IHSV | LLPKKL | ----- | AP | GKEGVGASLSQEF |

Figure 2.S6. Alignment of H2B sequences

Multiple sequence alignment of histone H2B from yeast, human and *Lingulodinium* is shown

| | | | | | | |
|------------|-------------------|----------------|------------------|-------------|-----------|-----------|
| | 1 | 10 | 20 | 30 | 40 | 50 |
| JO720817 | | PRQLPAAPPAPRN | PAGMALKEKKT | KMPPTADADGP | KKKIKK | KNTG |
| JO694219 | MEKGQK | EAMMADTKKA | KKKEKGKKIPKIGKAS | KRPKAPGVPM | RRGGRKKVE | |
| H2B_Hom_sa | | MPEPAKSAPAP | KKGSKKAVT | KAQKKDG | --- | KKRKRSRKE |
| H2B_Sac_ce | | MSSAAEKKPA | ASKAPAEKKPA | AKKTSTSV | --DG-- | KKRSKVRKE |
| | 60 | 70 | 80 | 90 | 100 | |
| JO720817 | SYATYIFKVLKQV | HPKMRISKQAMQIM | ESCVGDTFERLAG | EASRLSRMSK | | |
| JO694219 | TYAIYIYKVLKQI | HPECGISKRGMN | IMNSFMNDVFDRMA | TEATRLLRMTK | | |
| H2B_Hom_sa | SYSIYVYKVLKQV | HPDTCISSKAMG | IMNSFVNDIFERIA | EASRLAHYNK | | |
| H2B_Sac_ce | TYSYIYKVLKQT | HPDTCISQKSM | SILNSFVNDIFERIA | TEASKLAAYNK | | |
| | 110 | 120 | 130 | 140 | 145 | |
| JO720817 | RETLTSREIQSAVRLV | EPGELSKHAVSEG | CKAVMKFNAGNP | AADR | | |
| JO694219 | RRTLTSREIVETAVRLM | EPGELSKHAVSEG | TKAVTKYKQIG | | | |
| H2B_Hom_sa | RSITISREIQTAVRLL | EPGELAKHAVSEG | TKAVTKYTSSK | | | |
| H2B_Sac_ce | KSTISAREIQTAVRLI | EPGELAKHAVSEG | TRAVTKYSSSTQA | | | |

Figure 2.S7. Alignment of H3 sequences

Multiple sequence alignment of histone H3 from yeast, human and *Lingulodinium* is shown

| | | | | | | |
|-----------|-------------|---------------------------|-----------------|---------------|-----------|-----------|
| | 1 | 10 | 20 | 30 | 40 | 50 |
| JO753891 | MARGKQ | EAKRV | AGHECLRRKLDKN-- | RDGAML | TTCRDDGRV | RYKPGAKAL |
| JO722862 | MARTKQTARKS | AGGKAPRS | OLAAKAARKQ | TPVAGGIKKPH | - | RYRPGTVAL |
| JO740554 | MARTKQTARKS | AGGKAPRQT | LAAKAARKQ | APLLGGIKKPH | - | RYRPGTVAL |
| H3_Sac_ce | MARTKQTARKS | TGGKAPRKOLA | SKAARKS | SAPSTGGVKKPH | - | RYKPGTVAL |
| H3_Hom_sa | MARTKQTARKS | TGGKAPRKOLA | TKAARKS | SAPATGGVKKPH | - | RYRPGTVAL |
| | 60 | 70 | 80 | 90 | 100 | |
| JO753891 | QEIRHYQKTT | CLLINKISFORVVREL | CFKVQE | EMQLRPGKERALI | QEGEG | |
| JO722862 | REIRKYQKST | ELLIRKLPPFORLVREIAQDFKTDL | - | - | - | - |
| JO740554 | REIRKYQKST | ELLIRKLPPFORLVREIAQDFKTDL | - | - | - | - |
| H3_Sac_ce | REIRRFQKST | ELLIRKLPPFORLVREIAQDFKTDL | - | - | - | - |
| H3_Hom_sa | REIRRYQKST | ELLIRKLPPFORLVREIAQDFKTDL | - | - | - | - |
| | 110 | 120 | 130 | 140 | 150 | |
| JO753891 | MPAF | RFESQALLALQEA | EAYLVGLFEDANL | CAVHARRVTVM | PRDIQLSR | |
| JO722862 | --- | KFQSQAVLALQEA | EAYLVGLFEDTNL | CAIHAKRVTIM | PKDVQLAR | |
| JO740554 | --- | KFQSQAVLALQEA | EAYLVGLFEDTNL | CAIHAKRVTIM | PKDVQLAR | |
| H3_Sac_ce | --- | RFQSSAIGALQES | VEAYLVSLFEDTNL | AIHAKRVTIQ | KKDIKLAR | |
| H3_Hom_sa | --- | RFQSSAVMALQEA | SEAYLVGLFEDTNL | CAIHAKRVTIM | PKDIQLAR | |
| | 160 | | | | | |
| JO753891 | RIRGPEVLAG | | | | | |
| JO722862 | RLRGERI | | | | | |
| JO740554 | RLRGERI | | | | | |
| H3_Sac_ce | RLRGE | | | | | |
| H3_Hom_sa | RIRGERA | | | | | |

Figure 2.S8. Alignment of H4 sequences

Multiple sequence alignment of histone H4 from yeast, human and *Lingulodinium* is shown

| | | | | | | |
|-----------|---|-----|-----|-----|-----|----|
| | 1 | 10 | 20 | 30 | 40 | 50 |
| JO717937 | MSSVPGHSSSTKASKGDKFSKEA----- | | | | | |
| JO719134 | MAPSTPROVRPRPAAEVQPAAEELRPAARGARQPAVARPPPAALGPEVRATE | | | | | |
| H4_Hom_sa | MSSGRG----- | | | | | |
| H4_Sac_ce | MSSGRG----- | | | | | |
| | 60 | 70 | 80 | 90 | 100 | |
| JO717937 | ----- | | | | | |
| JO719134 | PEQLLRGLGVLACLAEGAPAAALGVLPALPSRPARLPSPRELRPALVAAVARR | | | | | |
| H4_Hom_sa | ----- | | | | | |
| H4_Sac_ce | ----- | | | | | |
| | 110 | 120 | 130 | 140 | 150 | |
| JO717937 | -----EKQP-VGKGGATRRHKVLRDNIQGITKPAIRRLARRGGVKKR | | | | | |
| JO719134 | LPRRADIAVAEGQPRLGKGGGKRSQKVIREHSAGITKSDLRRLARRAGCQR | | | | | |
| H4_Hom_sa | -----KGGKG LGKGGAKRRHKVLRDNIQGITKPAIRRLARRGGVKKR | | | | | |
| H4_Sac_ce | -----KGGKG LGKGGAKRRHKILRDNIQGITKPAIRRLARRGGVKKR | | | | | |
| | 160 | 170 | 180 | 190 | 200 | |
| JO717937 | ISGLIYEESSRGVLKTFLENVLRDSITYTEHARRKTVTALDIVYALKROGRT | | | | | |
| JO719134 | VAMLIYDEARAALASFLEKMLADITVYTEHTKRKTACPDVVLSLRRRGRV | | | | | |
| H4_Hom_sa | ISGLIYEEETRGVLKV FLENVIRDAVTYTEHAKRKTVTAMDVVYALKROGRT | | | | | |
| H4_Sac_ce | ISGLIYEEVRAVLKSFLESVIRDSVTYTEHAKRKTVTSLDVVYALKROGRT | | | | | |
| | 210 | | | | | |
| JO717937 | IYGFGL | | | | | |
| JO719134 | VYGA | | | | | |
| H4_Hom_sa | IYGFGG | | | | | |
| H4_Sac_ce | IYGFGG | | | | | |

2.6. Acknowledgements

Research support from the National Science and Engineering Research Council of Canada to DM (Grant number 171382-03) is gratefully acknowledged. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

CHAPTER 3 – PUBLICATION # 2

Transcripts from dinoflagellate tandem array genes are highly conserved and are not polycistronic

Mathieu Beauchemin^{*}, Sougata Roy^{*}, Philippe Daoust², Steve Dagenais-Bellefeuille, Thierry Bertomeu³, Louis Letourneau², B. Franz Lang⁴ and David Morse

Published in PNAS, 2012 109 (39): 15793-15798.

^{*}These authors have contributed equally to this work.

For this paper, I have contributed to designing the experiments and drafting some parts of the paper and I did the experiments required to generate the Figures 3.3A and B. I also did the extraction of RNA and performed the RACE reaction to verify by PCR the presence of SL in the 5' ends of potential HGT sequences. I contributed to the analysis required to generate the figures 3.S3 and 3.S5 and did all the analysis required to produce the Table 3.ST2.

3.1. Abstract

Dinoflagellates are an important component of the marine biota, but a large genome with high copy number (up to 5000) tandem gene arrays has made genomic sequencing problematic. More importantly, little is known about the expression and conservation of these unusual gene arrays. We assembled *de novo* a gene catalog of 74,655 contigs for the dinoflagellate *Lingulodinium polyedrum* from RNA-Seq (Illumina) reads. The catalog contains 93% of a *Lingulodinium* EST dataset deposited in GenBank and 94% of the enzymes in 16 primary metabolic KEGG pathways, indicating it is a good representation of the transcriptome. Analysis of the catalog shows a marked under-representation of DNA binding proteins and DNA binding domains compared to other algae. Despite this, we find no evidence to support the proposal of polycistronic transcription, including a marked under-representation of sequences corresponding to the intergenic spacers of two tandem array genes (TAG). We have also used RNA-Seq to assess the degree of sequence conservation in TAG and found their transcripts to be highly conserved. Interestingly, some of the sequences in the catalog have only bacterial homologs and are potential candidates for horizontal gene transfer. These were presumably transferred as single copy genes, and, since they are now all GC-rich, any derived from AT-rich contexts must have experienced extensive mutation. Our study has not only provided the most complete dinoflagellate gene catalog known to date, but has also exploited RNA-Seq to address fundamental issues in basic transcription mechanisms and sequence conservation in these algae.

Key Words: *De novo* assembly, *Lingulodinium polyedrum*, RNA-Seq, Transcription, Transcriptome

3.2. Introduction

Dinoflagellates are a group of freshwater and marine eukaryotes, and the marine photosynthetic species are important contributors to the ocean's primary production [131]. Dinoflagellates are notable for their symbioses with coral [432], the production of harmful algal blooms termed red tides [433], and their spectacular bioluminescence [434]. Dinoflagellates also display a number of unusual cytological and biochemical features. For example, their chromosomes remain permanently condensed throughout the cell cycle [404] and electron microscopy of dinoflagellate nuclei shows a whorled structure termed a cholesteric liquid crystal [116]. It seems likely that this unusual nuclear organization may impose restrictions on DNA replication and transcription, yet details of these processes are still unknown.

In part, study of dinoflagellate biochemistry is limited by a paucity of molecular tools, including the lack of a genome sequence and an inability to produce transgenic organisms [117]. One of the difficulties encountered is a generally large DNA content, with approximately 200 pg DNA (roughly 60 times that of the haploid human cell) reported for *Lingulodinium* [435]. Interestingly, two well-studied genes in this species are found in multiple copies arranged in tandem, peridinin-chlorophyll a-protein (PCP; ~5000 copies per nucleus) [159] and luciferase (146 copies) [158]. The presence of multiple gene copies is expected to render genome sequence assembly difficult, and, unless the different copies are well conserved, will also result in a complex transcriptome profile. Interestingly, the spacer regions in between the coding sequence of the tandem arranged genes have no recognizable transcription factor binding motifs, confounding attempts to understand how gene expression is regulated. This observation, in concert with the discovery of trans-splicing in dinoflagellates, has led to the proposal that dinoflagellate transcripts are polycistronic [266]. This proposal is largely derived from studies in kinetoplastids, where trans-splicing and polyadenylation are used to excise open reading frames (ORFs) from polycistronic transcripts [436]. While attractive, this proposal has not yet been tested experimentally in dinoflagellates.

Recent developments in high-throughput sequencing technologies have opened up an opportunity to examine transcriptomes of organisms as potentially gene-rich as the dinoflagellate *Lingulodinium*. We report here a transcriptome profile derived from Illumina sequencing, a technique commonly called RNA-Seq [437]. In addition to providing a measure of gene expression, the different sequences can be assembled *de novo* in order to develop a transcript profile. We have used RNA-Seq to obtain the most comprehensive gene catalog for any dinoflagellate described to date and to address fundamental issues in gene conservation and expression.

3.3. Materials and Methods

3.5.1. Cell culture

Unialgal but not axenic *Lingulodinium polyedrum* (CCMP 1936, previously *Gonyaulax polyedra*) was obtained from the National Center for Marine Algae (Boothbay Harbor, Maine). Clonal cell cultures derived from a single cell were grown in f/2 medium prepared from Instant Ocean under 12 h light (40 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ cool white fluorescent light) and 12 h dark at a temperature of $18 \pm 1^\circ \text{C}$ as described [438].

3.5.2. RNA purification and sequencing

Lingulodinium cultures were harvested at midday (LD 6) and midnight (LD18) under a light dark cycle, the corresponding times under constant light (LL 6 and LL 18), and from a culture taken at LD 18 but grown without added nitrate for four days. Cells were concentrated by centrifugation (500 x g for 1 minute), washed with fresh seawater and recentrifuged to reduce bacterial contamination. Total RNA was isolated by extracting cell pellets with Trizol (Invitrogen) and enriched for poly(A) RNA using the Poly(A)-Tract mRNA isolation system (Promega). RNA samples were subjected to quality control assessment using a Bioanalyzer (Agilent), and sequencing used an mRNA-Seq sample preparation kit (Illumina). Each sample was sequenced on a single lane of a Genome Analyzer IIX platform at the McGill University and Genome Quebec Innovation Center (Montreal, Quebec). In total, 312 M 76 base paired end reads (24 Gb total sequence) were obtained. RNA samples used for sequencing were also used for 5' and 3' RACE using the SMARTer RACE cDNA amplification kit (Clontech) using the manufacturer's protocol except that for 5' RACE, a version of splice-leader sequence modified to accommodate multiple splicing events [277] (5'-TGGCTCAAGCCATTTTGGCTCAAG-3') replaced the forward primer supplied in the kit.

3.5.3. Sequence assembly and analysis

After filtering sequences to exclude low quality bases, reads were assembled with Velvet and Oases. Hashlengths of 21-61 were tested, and kmers of 41 were retained for assembly. The original Velvet/Oases assembly resulted in 200,045 contigs of which 88,655

were >300 bp (the fragment size selected for sequencing). Some contigs contained more than one mRNA, possibly either due to alternative splicing or due to multiple gene copies with variation. To facilitate further analyses, only one transcript (best BLAST hit and/or longest sequence) was retained per contig to yield 74,337 contigs. In parallel, small scale assemblies (~3 M reads) were prepared from each sample using Geneious [439] at the default settings and screened against the Velvet/Oases assembly by BLASTn. These small-scale Geneious assemblies contained 318 generally high coverage transcripts absent from the Velvet/Oases assembly, and these were added back to produce a final dataset of 74,655 contigs >300 bp. All assembled sequences have been deposited in GenBank (Accession numbers JO692619 through JO767447).

Comparison of the catalog with known *L. polyedra* ESTs used a 65% GC-rich unigene dataset (BP742156-4266) [440] of which 56% had matches to other dinoflagellate ESTs (tBLASTn, e^{-10}). A second dataset also annotated as *L. polyedrum* (CD809360-810879) was not used as it was 54% GC-rich and only 42% had matches to other dinoflagellate ESTs. Sequence annotations and mapping to Gene Ontology (GO) [441] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [442] were performed using Blast2Go [443]. Interspecies comparison of GO and Protein Family (PFAM) domains [444] were made by similarly annotating *Paramecium tetraurelia* [352], *Thalassiosira pseudonana* [445], and *Chlamydomonas reinhardtii* [446] predicted gene models. Phylogenetic analysis was performed using the top 100 or 250 hits from BLAST [447] searches of the GenBank non-redundant (nr) database. Protein phylogenies were reconstructed using an online pipeline at Phylogeny.fr [402] that aligns the sequences using MUSCLE, curates them using GBlocks, performs phylogenetic analysis using PhyML then renders a tree using TreeDyn.

To detect genes of potential bacterial origin a three-step protocol was followed. First, candidates were selected by comparing the transcriptome an in-house bacterial protein database (prepared by downloading all available protein sequences from an NCBI database of whole genomes for 3 α -proteobacteria (*Rhodopseudomonas palustris*, *Sinnorhizobium sp*, *Azorhizobiumsp*), one β -proteobacteria (*Burkholderia dolosa*), two γ -proteobacteria

(*Pseudoalteromonas tunicata*, *Alteromonas macleodii*), one bacteroidetes (*Flavobacteriales bacterium*) and two cyanobacteria (*Prochlorococcus marinus*, *Synechococcus elongatus*) using BLASTx with E-values $< e^{-30}$. Next, any sequences with a match (tBLASTx, E-values $< e^{-30}$) to the *Karenia* or *Alexandrium* datasets were removed. In the last step, candidates were compared to the nr GenBank database using BLASTx to assess how many had matches only to bacteria in the top 100 sequences.

To assess sequence variation in the contigs, we obtained full-length *Lingulodinium* sequences from the nucleotide database at GenBank to use as reference sequences. These were used as scaffolds to assemble raw reads using the Geneious reference assembly function at the default settings [439] using a cutoff of 0.5% for SNP determinations. The number of synonymous and non-synonymous mutations was calculated as dS and dN by dividing by the total number of synonymous and non-synonymous positions in each sequence [448].

To find sequences common to several dinoflagellate species, EST datasets for *Karenia brevis* (66,657 sequences) and *Alexandrium sp* (50,302 sequences) were downloaded from GenBank, and each assembled separately using Geneious at default settings to produce unigene datasets containing 20,726 and 31,670 sequences for *Karenia* and *Alexandrium* respectively. The *Lingulodinium* dataset was compared to these datasets using BLASTx with E-values $< e^{-30}$.

3.4. Results

3.3.1. *The de novo assembly is an authentic portrait of the transcriptome*

RNA-Seq was used to generate 312 million sequence reads from a clonal *Lingulodinium* cell line. These reads were subsequently assembled to create a gene catalog containing 74,655 contigs of a length greater than 300 bp, a cut-off determined by the fragment size range selected for sequencing. The size distribution of the sequences in the catalog shows an exponential decrease in sequence number as a function of length (Supplementary Figure 3.S1.A), and while this is different from the size distribution of the poly(A) RNA (Supplementary Figure 3.S1.B), it is not unusual for *de novo* RNA-Seq assemblies [449].

The gene catalog is GC-rich, as expected [435], with a predominance of GC at the third codon position. To gauge the extent to which the catalog is an authentic representation of the transcriptome, we compared the gene catalog to a set of 2111 GC-rich *Lingulodinium* ESTs. Our catalog contains 93% of these ESTs with an average identity of 98% (Supplementary Figure 3.S2.). We also mapped the assembly on to sixteen primary metabolic KEGG pathways and find that the catalog contains 141 of the 150 enzymes expected (94%) (Supplementary Table 3.ST1.). Finally, we also tested the transcriptome for the presence of proteins involved in other basic metabolic processes such as DNA replication, transcription, mRNA transport, translation, splicing, ribosome biogenesis, and mRNA surveillance (Supplementary Table 3.ST2.). For this analysis, we prepared a dataset containing authentic KEGG component sequences for the different processes from whole genomes of mammals, plants, apicomplexans, ciliates and diatoms. This dataset was used to screen the *Lingulodinium* catalog (BLASTx E-value < e^{-20}). Overall, our transcriptome contains 64% of the proteins used by mammals, a value similar to what is found for the phylogenetically related apicomplexans (60% of mammalian sequences) and ciliates (63%). One notable exception is the absence of the TFIID subunit TATA-binding protein (TBP),

which is replaced in dinoflagellates by a TBP-like protein of different binding specificity [179]. We conclude our catalog is a good representation of the transcriptome.

Using an E-value cutoff of e^{-05} , tBLASTn analysis of the catalog showed that 25% of the 74,655 sequences had an annotated match, 45% had a non-annotated match, and 30% lacked similarity to any known sequence in GenBank. The annotated sequences in the catalog, when classified into gene ontology (GO) categories, show an under-representation of proteins classified as DNA binding compared to the ciliate *Paramecium*, the diatom *Thalassiosira* and the green alga *Chlamydomonas* (Figure 3.1.A). The gene catalog was also used to determine the number of matches to protein family (PFAM) DNA binding domains (Figure 3.1.B). *Lingulodinium* has representatives for only half the known DNA binding domains found in ciliates and diatoms, and completely lack the heat shock factor domains that comprise 0.8% of the diatom sequences. These domains present include the four core histone domains, a finding described elsewhere [189]. Importantly, the majority (68%) of all the *Lingulodinium* DNA binding domains fall into the class of cold shock domains. However, since these domains also bind mRNA, and are better known for their role in post-transcriptional regulation in eukaryotes [351], we conclude that *Lingulodinium* shows a marked under-representation in the types of proteins and protein domains involved in regulating transcription.

As a corollary to the underrepresentation of DNA binding factors and domains, we reasoned that if dinoflagellates generally favored post-transcriptional regulation mechanisms, components involved in these processes might be enriched in sequences shared between different dinoflagellate species. We obtained ESTs for *Karenia brevis* and for several *Alexandrium* species from GenBank, and aligned the sequences to produce datasets containing 20,726 *Karenia* unigenes and 31,670 *Alexandrium* unigenes. We then searched these unigene datasets with our *Lingulodinium* transcriptome using tBLASTn (E-value cutoff e^{-20}). The 5904 sequences shared among the three species were termed “core” dinoflagellate candidates (Supplementary Figure 3.S3.A). When compared to the *Lingulodinium* transcriptome, the core sequences are indeed depleted in DNA binding proteins and in addition, were enriched in proteins with kinase activity (Supplementary Figure 3.S3.B). The

ratio of core sequences to the total number in our catalog for different molecular functions shows a marked enrichment in translation factors, protein kinases and protein phosphatases, while DNA binding proteins are again underrepresented (Supplementary Figure 3.S3.C). This analysis supports the contention that dinoflagellates may preferentially regulate gene expression using post-transcriptional control mechanisms.

3.3.2. Tandem gene array sequences are highly conserved in the transcriptome

To assess the degree of sequence conservation in transcripts from the tandem array genes, raw reads were assembled to a reference gene (mean coverage $\sim 10,000$) and the number of variant nucleotides at each position measured after trimming to remove low quality bases. For PCP, (~ 5000 copies), previous work has indicated that both the coding sequences and intergenic spacers are highly conserved [159], and this was confirmed in the transcript sequences at high coverage (Figure 3.2.A). When the entire PCP coding region is scanned nucleotide by nucleotide, few positions show a level of sequence variation greater than 0.5% (dotted line in Fig 3.2.A). To quantify the degree of sequence conservation, we counted the number of positions with different levels of variation (Supplementary Figure 3.S4.A). This spectrum of variation shows that over half of the positions have a level of variation corresponding to the Q30 (99.9% accuracy) of the sequencing reaction. This level of sequence conservation is similar to that observed for ribosomal RNA transcripts (Supplementary Figure 3.S4.B).

To verify if there was a systematic bias toward synonymous mutations, raw reads were assembled onto several nuclear-encoded and plastid-encoded reference gene scaffolds (average coverage > 1000). All positions with greater than 0.5% variant nucleotides were classified as synonymous or non-synonymous and normalized to the total number of synonymous or non-synonymous positions in the sequence (dS and dN) (Figure 3.2.B). Curiously, while the dN/dS ratio is close to one for the plastid-encoded genes (open circles), nuclear-encoded genes (closed circles) show a clear tendency toward synonymous changes, indicative of a purifying selection. It is also of interest that the two sequences with the largest number of synonymous mutations, rubisco and glyceraldehyde-3-phosphate dehydrogenase sequences, are both thought to be derived from horizontal gene transfer (HGT) [450, 451].

3.3.3. Sequences of potential bacterial origin have the same GC-content as the host

To pursue the possibility that sequences derived from HGT might act as a counterpoint to the high degree of sequence conservation observed for tandem array genes, we searched our catalog for suitable HGT candidates. We reasoned that sequences derived from HGT were likely to have been transferred as single copy genes and thus may not have been subject to the same sequence conservation mechanisms. A previous analysis of the dinoflagellate *Karenia* has shown that 2.4% of the genes were of potential bacterial origin as determined by “best hit” BLAST searches, and 0.3% were uniquely found in bacteria and dinoflagellates [45]. We also used these criteria to try and uncover examples of HGT in our catalog. We first compared the catalog to an in house bacterial protein database using BLASTx, a search that returned 2354 sequences (~3%). Interestingly, similar searches using our *Karenia* and *Alexandrium* unigene datasets against the same bacterial protein databank yielded a similar fraction, and all the sequences recovered by the search have the same average GC-content as the ensemble of sequences in the species from which they are derived (Figure 3.3.A). We also note that of the 2354 *Lingulodinium* sequences, the majority (~80%) are targeted to the cytoplasm or to other cellular organelles (Supplementary Figure 3.S5.A), indicating only a minority are likely to be mitochondrial or plastid components transferred to the nucleus from the endosymbiont [452]. Most are enzymes and the domain structure shows enrichment in nucleotide binding or biosynthetic functions (Supplementary Figure 3.S5. B, C).

Among these 2354 sequences, we next determined candidates for HGT following divergence of *Lingulodinium* by removing any sequences also found in the *Karenia* and *Alexandrium* datasets. This reduced the number of potential HGT to 1422 sequences. Lastly, we tested if any of these were found uniquely in bacteria and dinoflagellates using BLASTx to screen the GenBank nr dataset. As our goal was to find potential examples of HGT, rather than to determine the full extent of HGT contribution to the *Lingulodinium* genome, only a 200 sequence subset of the 1422 potential HGT sequences were tested, and of these, 58 sequences returned only bacterial homologs in the top 100 BLAST hits.

To assess the possibility that the transcriptome might contain sequences resulting from bacterial contamination of our unialgal but not axenic cultures, we examined the initial 2354 sequences for the presence of the characteristic 22-nucleotide *trans*-spliced leader (SL) sequence. The presence of a 5' SL sequence constitutes an unambiguous marker for dinoflagellate nuclear transcripts. Bioinformatics searches found that 60 of the 2354 putative bacterial sequences (2.5%) had a partial (10 nucleotide) match to the SL, while 1420 sequences in the full catalog (2%) contained the same partial match. These percentages are low presumably because read coverage is always low at the ends of our contigs and because many of the contigs are small and are likely to be only fragments of longer sequences. We also tested for the presence of the SL directly using a 5' RACE reaction between an SL 5' primer and a sequence-specific 3' primer. Here we used 14 random sequences of the 58 found only in bacteria and *Lingulodinium*. Two of the 14 sequences, highly abundant in the transcriptome as based on read counts, and two of an additional 12 low abundance transcripts were successfully amplified. By comparison, out of 6 non-bacterial sequences, four highly abundant sequences were amplified while two low abundance transcripts were not. We attribute the difficulty in amplifying low abundance transcripts to the fact that the SL primer will bind to all transcripts, lowering its effective concentration and selectively disadvantaging amplification of low abundance transcripts.

We then analyzed the phylogeny of the four *Lingulodinium* sequences that both contained an SL by 5' RACE and had only bacterial homologs. As exemplified by the arabinofuranosidase gene (JO761275), three had AT-rich bacteria as their closest phylogenetic neighbors (Figure 3.3.B), while the fourth could not be assigned to a particular clade. Since any sequence derived from an AT-rich organism after divergence of *Lingulodinium* and *Alexandrium* has altered its GC-rich content, this indicates single-copy genes may not be subject to the same degree of sequence conservation as are TAG.

3.3.4. Assessing the potential for polycistronic transcripts

Transcription in dinoflagellates is poorly understood, and the discovery of a trans-spliced leader at the 5' end of all dinoflagellate transcripts [266], in conjunction with the unusual tandem arrangement of gene copies and the lack of recognizable promoter elements in the intergenic spacer [159], has led to the proposal that dinoflagellates may synthesize long polycistronic transcripts [453]. In this model, mature transcripts from tandem array genes have a single origin of transcription and the individual ORFs are excised by trans-splicing at their 5' end and by cleavage followed by polyadenylation at the 3' end. Interestingly, the model makes several predictions concerning the amount and type of RNA in the transcriptome that can be tested by deep sequencing.

One prediction is that reads corresponding to the genomic sequence between coding sequences, termed intergenic spacer regions, should be present in the RNA. However, when reads are assembled to the genomic sequence of PCP, few reads are found corresponding to the spacer (Figure 3.4.A, B), and this is also found for the luciferase TAG (Supplementary Figure 3.S6.A, B). As a control, since there are no known introns in *Lingulodinium*, we used the polycistronic ribosomal RNA precursor transcript which contains two internal transcribed spacers (ITS) that are excised during processing. The RPKM values for mature rRNAs are 7 to 20 times greater than read counts for ITS1 and ITS2, respectively. In contrast, RPKM values for the coding sequences of luciferase and PCP are 5000 to 36000 times greater than for their respective spacer regions (Figure 3.4.C). We conclude that non-coding RNAs from tandem array genes do not accumulate to an appreciable extent in the transcriptome.

A second prediction is that the number of mature transcripts should be roughly proportional to the number of gene copies. To test this, we counted the number of reads assembled to the reference sequences for the five genes for which gene copy numbers are known and counted reads per kilobase of transcript per million reads (RPKM [454]). This analysis (Figure 3.4.D) shows not only that genes with different copy number can have

similar transcript levels (compare PKA with cyclin) but also that genes with the similar copy number can have different abundance in the transcriptome (compare PCP and cyclin).

A last prediction is that a polycistronic transcript processed into its repeat units should produce equal numbers of each ORF. To test this, we examined the sequence variation in luciferase gene transcripts (Supplementary Figure 3.4.C). This gene has 146 copies, and if transcripts from each gene accumulated to the same extent, any position where only one gene has a mutation would result in a sequence variation of $1/146$ ($\sim 0.7\%$). Thus, if many positions were mutated in only one out of the 146 copies, our analysis would show a peak at 0.7% (simulated in Supplementary Figure 3.S4.D). Since a peak at 0.7% is not observed in our data, we conclude that transcripts from all 146 gene copies do not accumulate equally in the transcriptome.

3.5. Discussion

We have used RNA-Seq to profile the transcriptome of the dinoflagellate *Lingulodinium* with an initial aim of providing a gene catalog to facilitate gene discovery. *De novo* assembly is difficult, so to evaluate the quality and completeness of our transcriptome, we determined the coverage of *Lingulodinium* ESTs in GenBank and of 16 different primary metabolic KEGG pathways. This indicates that 93-94% of the transcriptome is represented in our catalog.

A global GO analysis of the catalog reveals some striking features about the types of genes present. Of particular interest are the DNA binding proteins, which unless they are very different in the dinoflagellates, are remarkably under-represented. Little is known about regulation of gene expression in dinoflagellates, but this observation suggests that transcriptional control may not be used as extensively as is the case for other organisms. Interestingly, the most abundant DNA-binding domain, the cold shock protein domain, is also associated with post-transcriptional regulation in eukaryotes [351] and thus may not function as a DNA-binding protein at all in *Lingulodinium*. The idea that dinoflagellates favor regulation of gene expression at a post-transcriptional level agrees with studies on circadian regulation of protein synthesis showing extensive translation control [77, 79]. It is also interesting that while our transcriptome contains all four core histones as well as a suite of histone modifying enzymes, the histone RNA levels are low compared to higher plants, and the histone proteins are still below the level of detection using antibodies [189]. Thus, dinoflagellates may not have extensive access to modified histones as a means of regulating transcription rates.

The possibility that very low levels of histones are present in the dinoflagellates is intriguing, as low levels of acetylated histone H3 are used to initiate polycistronic transcription in the kinetoplastids [214]. Kinetoplastids transcribe a polycistronic RNA in both directions from a central point on the chromosome, then excise the individual ORFs by addition of a trans-spliced leader [455]. This similarity with the trans-splicing of dinoflagellate transcripts [266] has led to the proposal that polycistronic transcription of

tandem array genes (TAG) might occur in dinoflagellates. Here, TAG would be transcribed as a single transcript with multiple ORFs, and processed by trans-splicing and polyadenylation to yield equal numbers of all individual ORFs. We have tested this model experimentally using our RNA-Seq data. First, we examined the raw RNA read data for sequences that could be assembled to the intergenic spacer sequences, reasoning that polycistronic transcripts should produce intergenic regions and coding sequences in initially equal amounts. We anticipated these non-coding sequences might, like introns, be easily detectable, as intron sequences accumulate to roughly 1% of sequence reads in fission yeast [456] and are even more abundant in mammals [457]. Since there are as yet no known introns in *Lingulodinium*, we instead used read counts corresponding to the ITS, a spacer region excised during formation of mature rRNA, and we find that the ITS to mature rRNA read ratio is much higher than the read ratio of TAG spacer to mature ORF (Figure 3.4.C). One factor that may influence the spacer/coding sequence read ratio is a preferential loss of non-polyadenylated spacer regions during poly(A) purification. However, poly(A) RNA is only enriched about ten-fold in our preparations, clearly insufficient to account for the differences observed. We also tested for a correlation between the number of genes in the tandem gene array and the amount of transcript, which would be expected if TAG were in a single operon (or several co-regulated operons) and different transcripts were not differentially degraded. However, we find that transcript abundance does not directly correlate with copy number (Figure 3.4.D). Lastly, we also devised a test to determine if all of the genes in the 146 gene luciferase tandem array were equally expressed. We predicted that since a mutation in one of the copies in the array would have a sequence variation (0.7%), if a sizable number of positions were different in one of the 146 gene copies, we would see a peak of nucleotide variation at 0.7% (Supplementary Figure 3.S4.D). However, this predicted peak of variation was not seen in our data (Supplementary Figure 3.S4.C). Taken together, we conclude there is no support for the existence of polycistronic transcripts in *Lingulodinium*.

Interestingly, TAG transcripts are remarkably well conserved. This sequence conservation is seen at the nucleotide level, as the variation at each position of the gene sequence is low. This may suggest that, as for ribosomal genes, a tandem gene arrangement

may be conducive to the conservation of sequence by gene conversion. Indeed, the levels of sequence variation found in the transcriptome for PCP (~5000 copies) are similar to those observed for rRNA. It is important to note that this variation is contained within the raw reads, and thus a new assembly of reads to a reference sequence appears to be the only means of recovering all the variant nucleotides in their correct proportions. We also observe extensive sequence conservation at the deduced protein level, as variations in nuclear-encoded genes appear to be biased toward synonymous mutations. It is an intriguing question how mutations that lead to deleterious changes in the amino acid sequence of the protein might be traced back to the gene that encodes them in the context of a large gene family. It is possible that purifying selection might operate against a deleterious mutation only if it becomes fixed in the gene array by gene conversion.

A TAG arrangement is not by itself sufficient to confer a high degree of sequence conservation, as considerable sequence diversity has been observed for TAG in other dinoflagellates [18, 162, 278, 349, 458, 459]. This may involve the number of gene copies, as the ~36 PCP copies in *Symbiodinium* have multiple non-synonymous mutations in the CDS [459] in contrast to the ~5000 almost identical PCP copies in *L. polyedrum* (Figure 3.2.). Furthermore, the proximity of the elements in the TAG also seems important, as actin copies in *Amphidinium* are found in two separate genomic clusters with different nucleotide sequence, intron length and intergenic spacer size [162]. This suggests concerted evolution is allowed within TAG clusters but not between two different clusters of the same gene.

As a contrast to the sequence conservation observed for TAG, we sought genes that may at one time have been low or single copy and thus may have been allowed to mutate. To this end we searched the transcriptome for sequences potentially derived from bacteria, since many marine bacteria are AT-rich, and because HGT would have placed these sequences in a GC-rich environment. We reasoned that if any of these genes were originally AT-rich but are now GC-rich, *Lingulodinium* must be able to extensively modify the sequence of single copy genes. We found several examples of sequences with AT-rich phylogenetic relatives and an unambiguous dinoflagellate origin based on presence of the SL. Interestingly, since roughly a third of these quences with similarity to bacterial proteins are found with matches to only

bacterial sequences in the top 100 BLAST hits, up to ~0.6% of *Lingulodinium* sequences may have a potential bacterial origin. While substantially lower than the 7.5% of sequences with a bacterial origin reported for the diatom *Phaeodactylum tricornutum* [460], this is clearly an interesting avenue to pursue in future work.

Our RNA-Seq derived gene catalog contains 74,655 unique sequences that would agree well with gene content estimates based on extrapolations of genome sizes [101] if all genes were present in a single copy. However, the smallest gene family known so far, Protein Kinase A, has 30 copies [461] and other genes have even higher copy numbers. In order to accommodate 75,000 genes of 1 kbp with an average gene copy number of 30 within the 200 pg nuclear DNA ($\sim 2 \times 10^{11}$ bp) we would have to assume a gene density of 1%. So far, the only report of a large genomic DNA fragment sequence (230 kb) is in *Heterocapsa* and this indicates a gene density of only 0.2% [462]. Furthermore, the size distribution of our sequences, biased toward short sequences (Supplementary Figure 3.S2.), also suggests the total number of genes will be less than ~75,000.

We report here the most extensive transcriptome profile yet presented for a dinoflagellate, and our analysis of the gene catalog suggests dinoflagellates may favor post-transcriptional regulation of gene expression. We have also used the read data to explore the nature of tandem array genes and the mechanisms used for their expression. In particular we find no evidence for the polycistronic transcripts that are found in kinetoplastids, another organism with rampant *trans*-splicing. It appears that unraveling the mechanism of transcription in dinoflagellates will require extensive mining of data banks such as our *Lingulodinium* transcriptome as well as biochemical analyses to provide functional tests for DNA binding activities.

Figure 3.1. Global analysis of the *Lingulodinium* assembly

(A) Gene ontology of annotated sequences in the transcriptome show a decreased level of DNA binding proteins and an increased level of substrate specific membrane transporters compared to the ciliate *Paramecium*, the diatom *Thalassiosira* and the green alga *Chlamydomonas*. (B) The number of protein family DNA binding domains detected in *Lingulodinium* compared with those detected in *Paramecium*, *Thalassiosira* and *Chlamydomonas*. DNA binding domains not present in *Lingulodinium* (CCAAT, E2F, GATA, Helix-hairpin-helix, Helix-loop-helix, CBF, KN, HSF, Sigma-70, TAZ, CXC and WRKY) are not included. The groups are CBF-Core-binding factor; HSF-Heat shock factor; TAZ -Transcriptional coactivator with PDZ-binding motif.

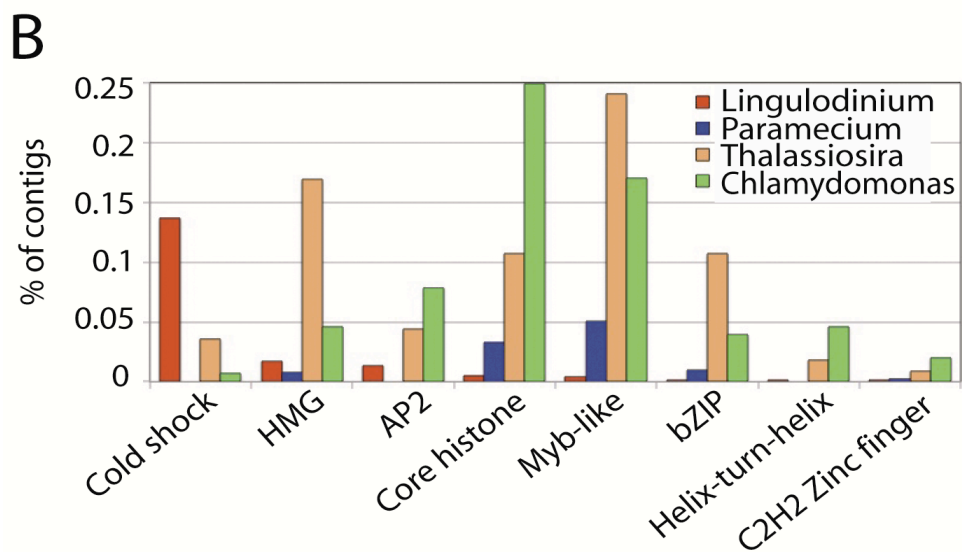
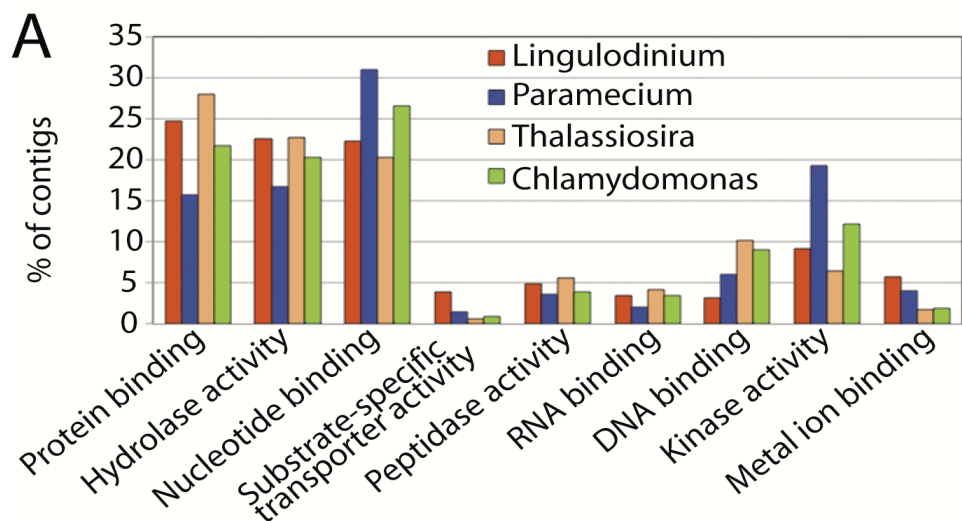


Figure 3.2. Sequence variation among transcripts

(A) The nucleotide variation (% total reads with a nucleotide different from a peridinin-chlorophyll a-protein (PCP) reference sequence at each position) is given along the PCP sequence. The dotted horizontal line at 0.5% variation is the threshold used for calculating dS and dN. (B) The ratio of non-synonymous (dN) to synonymous (dS) changes is shown for NCBI reference sequences with greater than 1000-fold coverage. The dotted line ($dN/dS = 1$) represents neutral selection. Plastid-encoded (open circles) and nuclear-encoded (closed circles) sequences are shown separately. The positions of Rubisco (RUB) and Glyceraldehyde-3-phosphate dehydrogenase (GAP) are indicated.

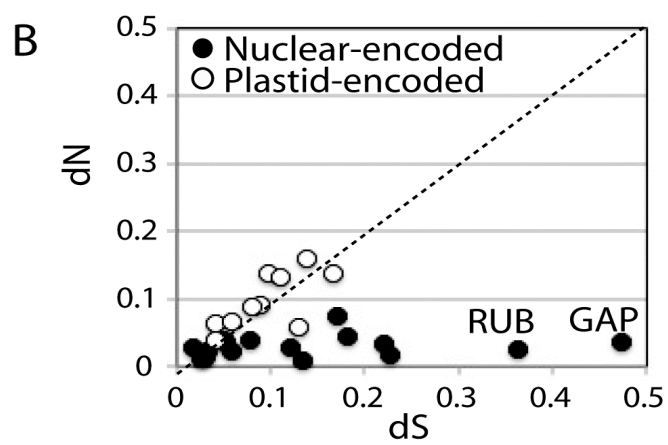
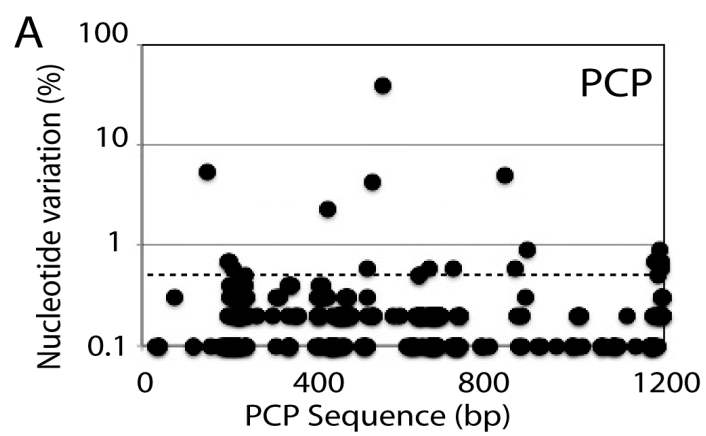


Figure 3.3. Bacteria-like sequences in the transcriptomes of different dinoflagellates have GC-contents commensurate with the host

(A) The GC-content of the unigene catalogs as well as for potential bacterial sequences was compared for data sets from *Lingulodinium* and two other dinoflagellate species, *Alexandrium sp* and *Karenia brevis*. A similar proportion of potential bacterial sequences (3%) are found in all three datasets. Solid lines represent the entire dataset, dotted lines the dataset of potential bacterial sequences. (B) The GC-content of an arabinofuranosidase (JO761275) in *Lingulodinium polyedra* (Linpo) is higher than the more closely related eubacterial sequences (*Butyrivibrio fibrisolvens*, *Herpetosiphon auranticus*, *Bacillus clausii*, *Oceanobacillus iheyensis*, *Clostridium sp*, an unidentified γ -Proteobacteria and *Flavobacteriales sp*) and is more similar to the more distantly related archeal sequences (*Halorhabdus utahensis* and *H. tiamatea*).

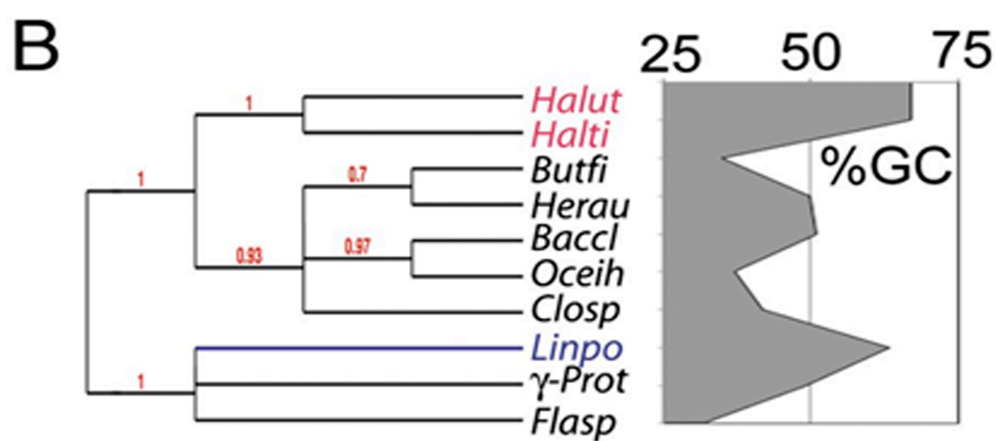
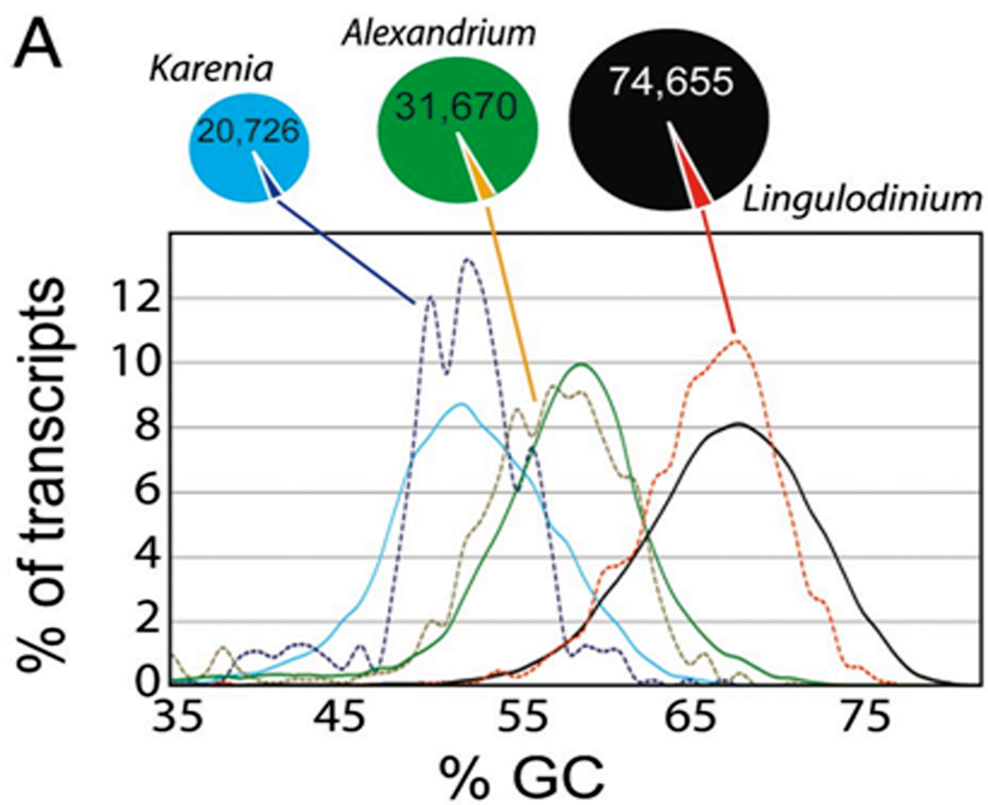
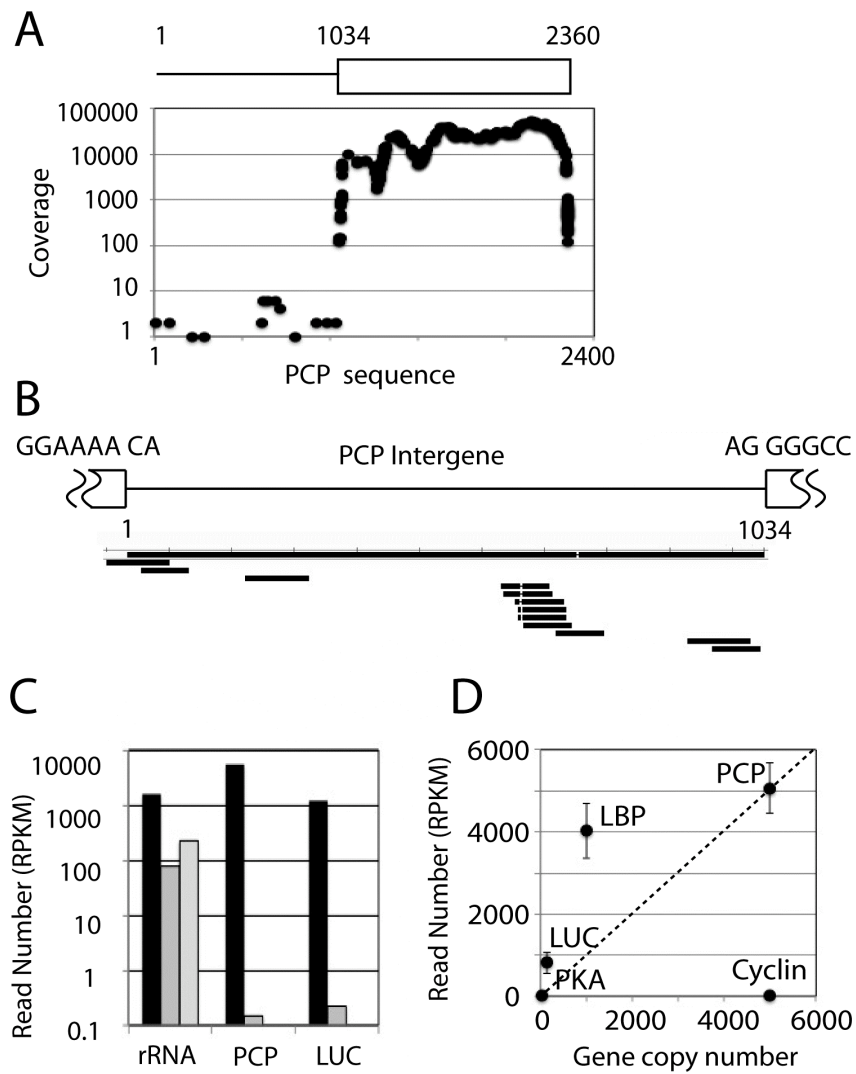


Figure 3.4. RNA-Seq does not support a polycistronic transcription mechanism

(A) A PCP genomic sequence (Accession number GPU93077) containing the intergenic spacer (line) and coding sequence (box) was used to align 0.58M reads from a dataset of 89M reads. The coverage (number of reads) is shown at each position. (B) Detail of the read assembly to the 1034 bp intergenic spacer region. Sequences corresponding to the polyadenylation site (left) and trans-splice site (right) were determined by comparing the assembly with the genomic sequence. (C) Individual reads (as reads per kilobase per million, RPKM [454]) aligning to the mature rRNA (dark bars) and each of the two internal transcribed spacers (gray bars) compared to both luciferase and PCP coding sequences (dark bars) and intergenic spacers (gray bars). (D) Read counts (mean \pm SD of four independent samples) plotted as a function of gene copy number for genes with known copy number (luciferin binding protein, LBP; peridinin chlorophyll a protein, PCP; Protein Kinase A, PKA; luciferase, LUC).



Supplementary Figure legends

Figure 3.S1. Size distribution of sequences in the transcriptome and in the mRNA

(A) The size distribution plotted as a histogram of number of sequences for each contig length in the 74655 sequence transcriptome. (B) The size distribution of the RNA sample used for sequencing as determined using a Bioanalyzer. The sizes of the two peaks of ribosomal RNA still visible in the analysis are shown.

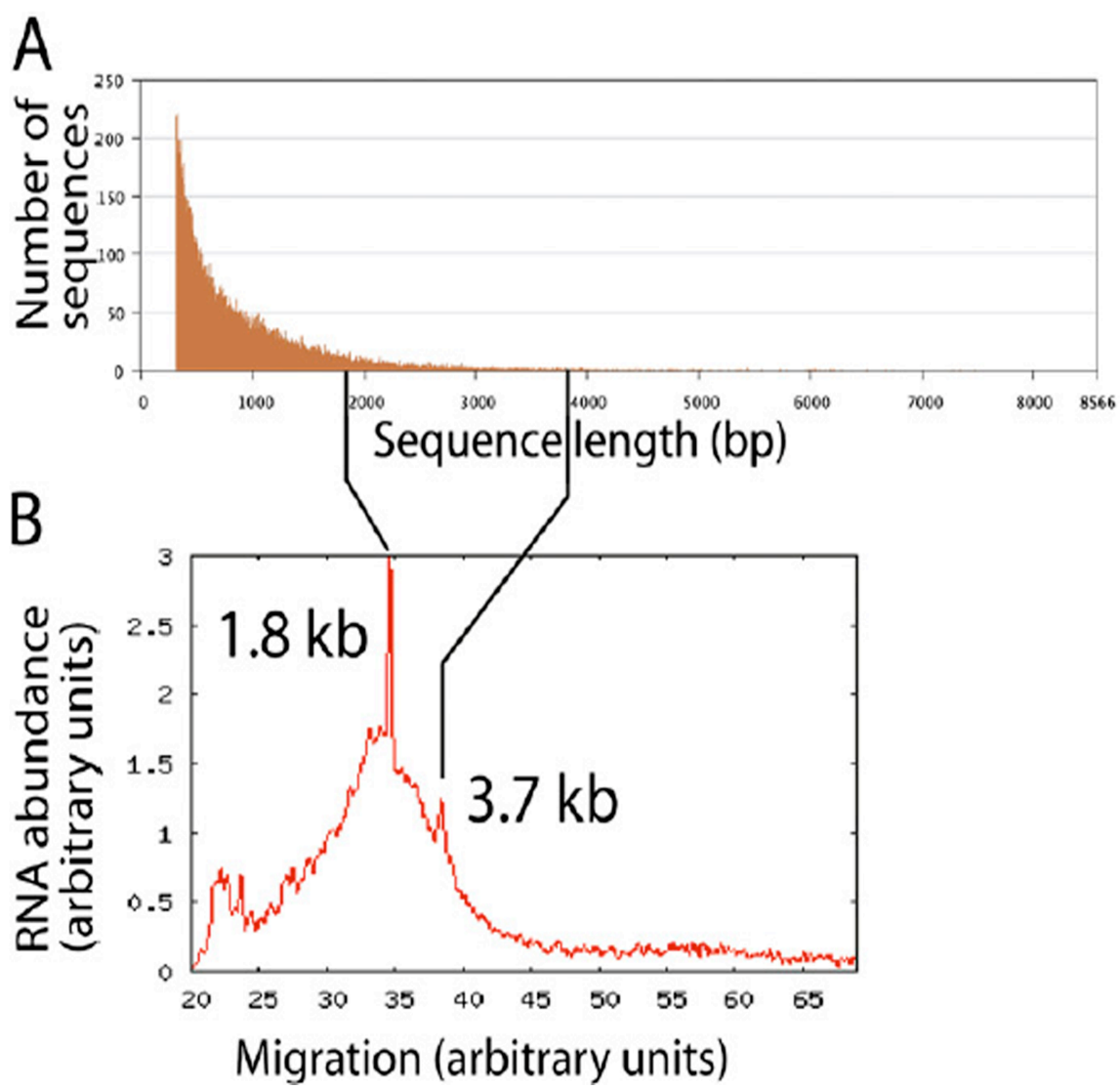


Figure 3.S2. Degree of sequence identity of *Lingulodinium* ESTs with the transcriptome

The degree of sequence identity as a function of the proportion of the EST sequence covered is shown by a comparison of the transcriptome sequences with 2111 GC-rich *Lingulodinium* ESTs in GenBank. Each point represents a Sanger EST with a corresponding sequence in the transcriptome. Due to the short average length of the transcriptome sequences, there many ESTs that are incompletely covered by the transcriptome contigs, and several ESTs have matches with more than one contig.

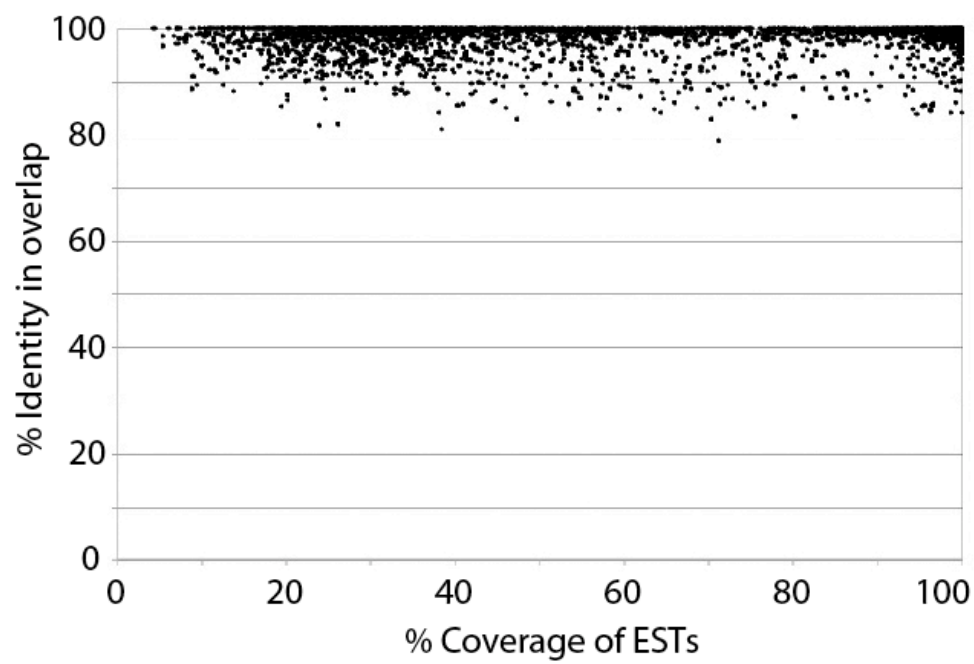
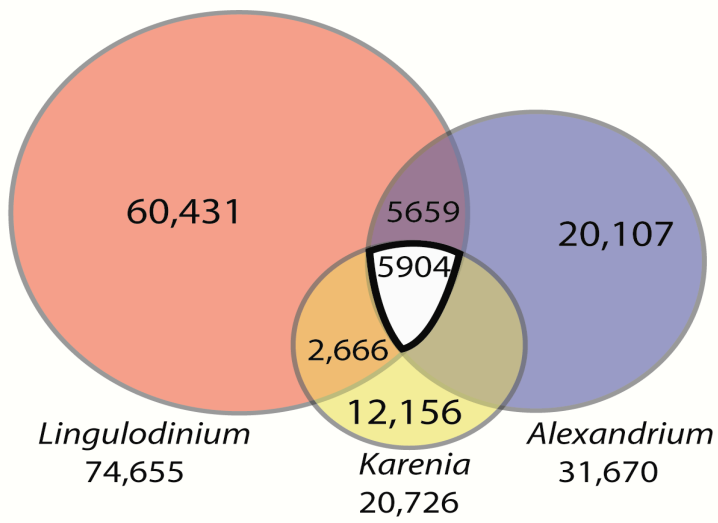


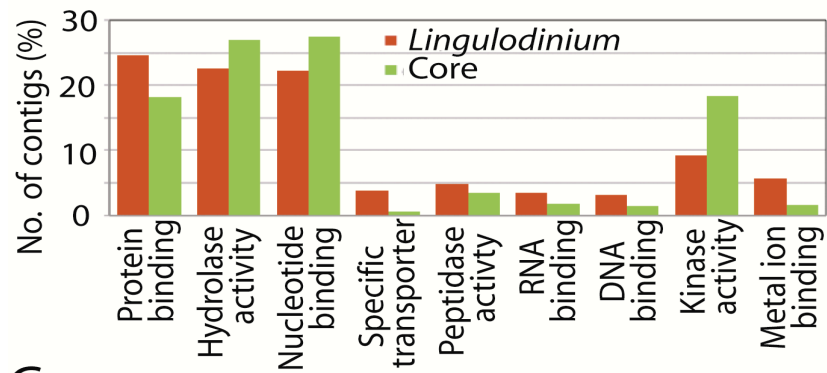
Figure 3.S3. Characterization of sequence common to *Lingulodinium*, *Alexandrium* and *Karenia*

(A) The transcriptome was mapped onto *Alexandrium* and *Karenia* EST unigenes using tBLASTn at a cutoff of e^{-20} . (B) Compared to the full *Lingulodinium* transcriptome, the “core” dinoflagellate sequences are enriched in the kinase activity category by gene ontology classification. (C) The ratio between the number of sequences in the “core” dinoflagellate dataset and the full transcriptome was calculated for each gene ontology molecular function classification. Groups found enriched in the “core” dataset are shown in green, and include sequences involved in translation and post-translation control. The Blast2Go was used with the default parameters for blastX and further annotation into different categories.

A



B



C

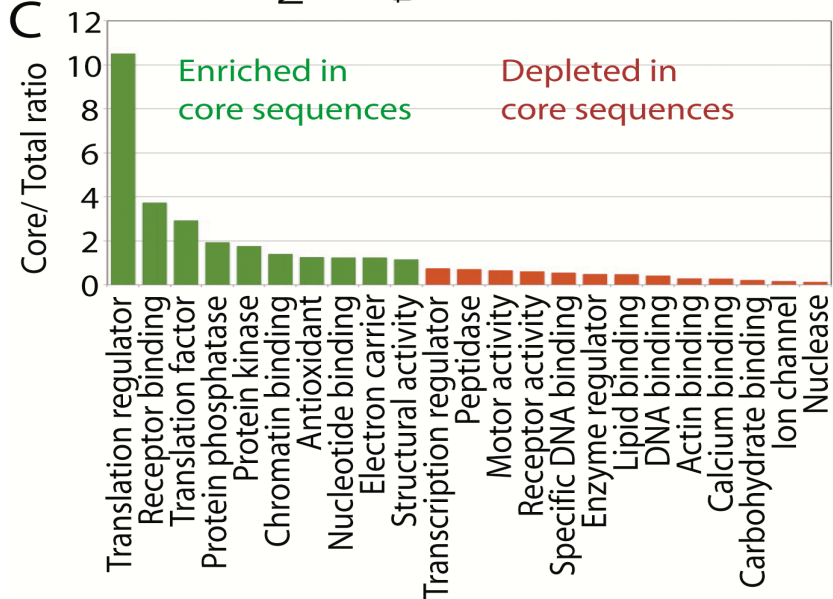


Figure 3.S4. Frequency spectrum of sequence variation in PCP and Luc TAG transcripts

The sequences of PCP (A), rRNA (B) and Luciferase (C) were first used to align readstrimmed to remove low quality or ambiguous bases from a dataset containing 89 million (M) reads. A plot of this data directly displays the nucleotide variation at each position (as in Figure 2A for PCP). This data is then used to determine the frequency spectrum of polymorphic variation by counting the number of times each percent nucleotide variation is observed over the sequence. Bin sizes are 0.1% up to 1%, between 1% to 10%, and 10% up to 100%, and the data is reported as the fraction of total sites with a given level of variation. The dotted vertical line represents the same 0.5% variation shown in Figure 2A, and the high number of sites with a sequence variation of up to 0.1% reflects at least in part error in the sequencing reaction (a Q30 corresponds to an accuracy of 99.9%). The peak at 1% variation in (B) reflects an increased bin size (from 0.1% to 1%). (D) A hypothetical frequency spectrum of polymorphic variation in luciferase gene transcripts was constructed by arbitrarily adding 100 nucleotide positions with a 0.7% variation to the data prior to binning. The predicted peak at 0.7% variation observed in the frequency spectrum corresponds to what would be expected if one of the 146 gene copies differed from the others at 100 of the 4000 positions in the gene sequence and transcripts from all gene copies accumulated equally in the transcriptome.

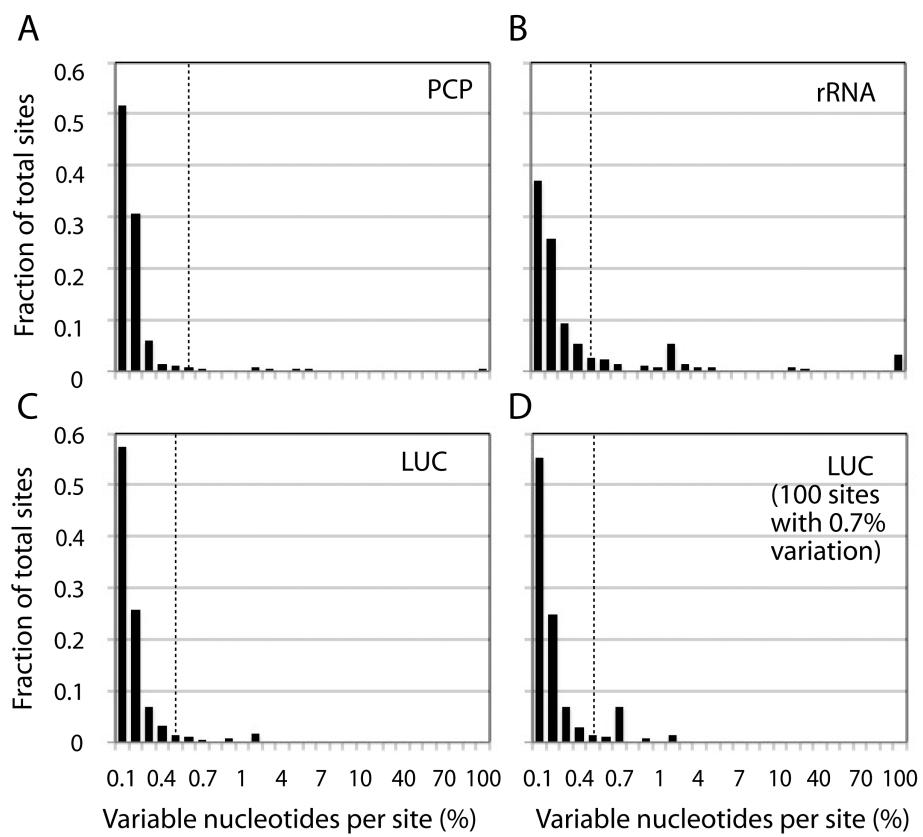


Figure 3.S5. Characteristics of the bacterial-like sequence in the transcriptome

(A) 2354 sequences were identified as putative bacterial sequences and 414 of these had an annotated match to GenBank. The annotated sequences were classified into the gene ontology compartment categories cytoplasm, mitochondria, chloroplast or all other membrane bound compartments. (B) Functional classification shows enrichment of sequence in nucleotide binding and enzymatic GO categories. (C) The 30 most abundant PFAM domains include principally nucleotide binding and biosynthetic functions.

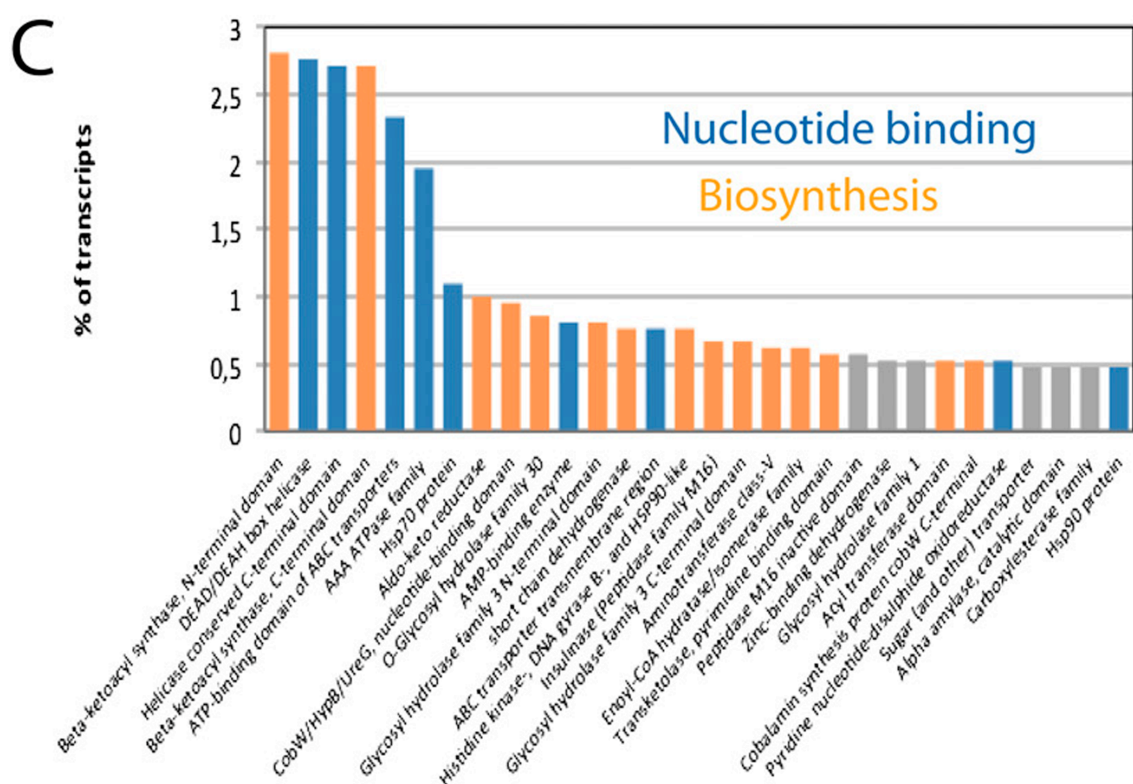
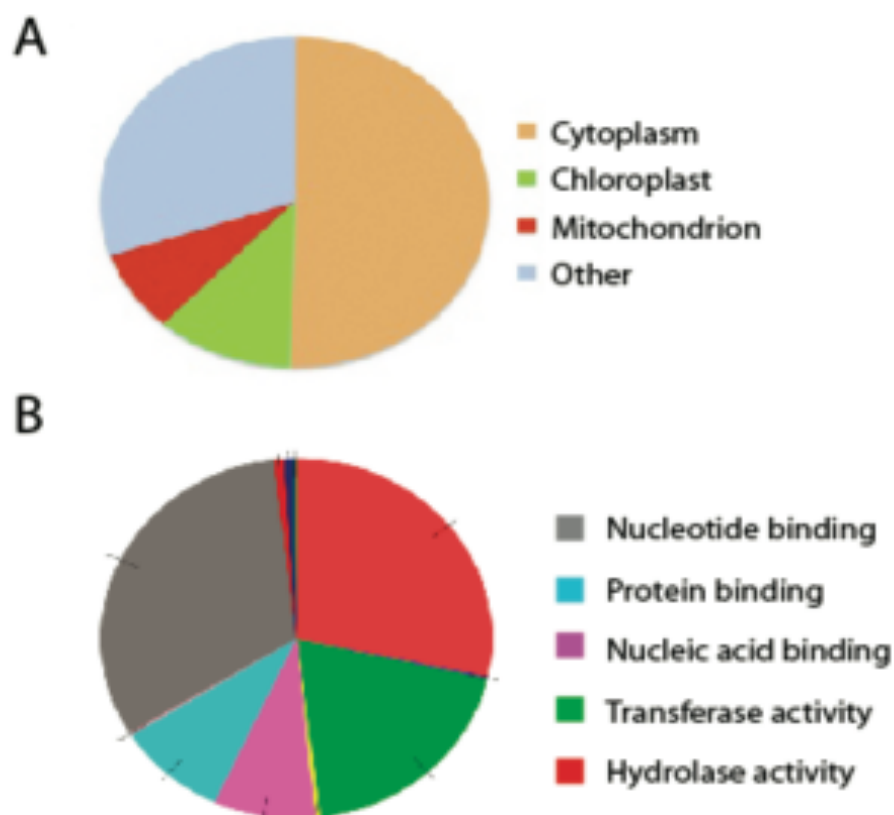
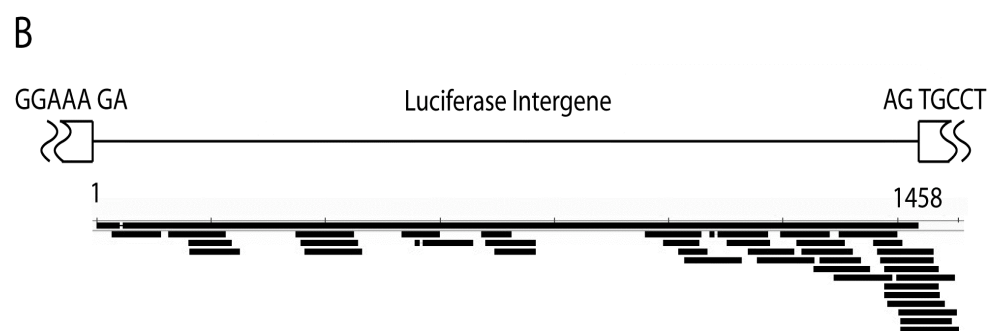
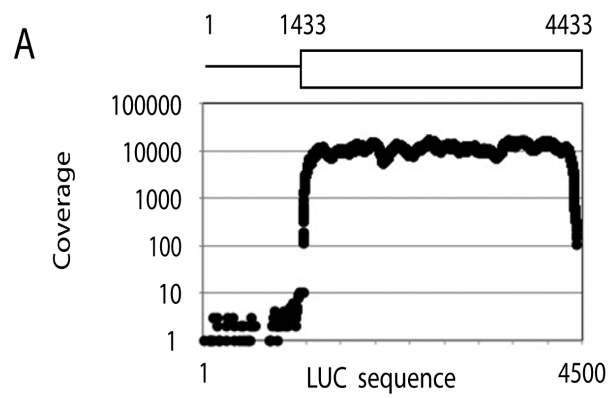


Figure 3.S6. Detection of reads aligning to the Luciferase intergenic spacer

The genomic sequence of the Luciferase tandem array unit was used to align reads trimmed to remove low quality or ambiguous bases from a dataset containing 89 million (M) reads. The alignment of 0.43 M reads with the intergenic spacer (line) and coding region (box) is displayed as coverage (number of reads) at each position (A). A detail of the reads assembled only to the intergenic spacer is shown (B) together with the nucleic acid sequences at the site of trans-splicing (right) and the polyadenylation site (left).



Supplementary Tables

Table 3.ST1. Number of KEGG genes found for a variety of pathways

| Pathway | <u>Essential Enzymes</u> | | Worst |
|--------------------------------------|--------------------------|------------|------------|
| | Expected | Found | e-Value |
| Glycolysis ⁽¹⁾ | 10 | 10 | e^{-87} |
| TCA cycle | 9 | 9 | e^{-107} |
| Oxid. Phosphorylation ⁽²⁾ | 5 | 5 | e^{-102} |
| Carbon Fixation | 11 | 11 | e^{-71} |
| Purine Synthesis | 18 | 17 | e^{-56} |
| Pyrimidine synthesis | 12 | 11 | e^{-58} |
| Fatty Acid synthesis | 6 | 4 | e^{-84} |
| Fatty acid oxidation | 4 | 4 | e^{-138} |
| F, Y, W synthesis | 16 | 15 | e^{-50} |
| S, G, T synthesis | 9 | 9 | e^{-64} |
| R, P synthesis | 11 | 11 | e^{-56} |
| A, D, N, E, Q synthesis | 5 | 5 | e^{-118} |
| C, M synthesis | 9 | 8 | e^{-117} |
| V, L, I synthesis | 8 | 7 | e^{-97} |
| K synthesis | 9 | 6 | e^{-94} |
| <u>H synthesis</u> | 9 | 9 | e^{-55} |
| <u>Total</u> | <u>151</u> | <u>141</u> | |

⁽¹⁾ Glucokinase replaces hexokinase in the first reaction.

⁽²⁾ Only the core proteins of the different complexes were analyzed. Most of the subunits in each of the five complexes were well represented.

Table 3.ST2. Number of KEGG pathway sequences found in mammals, plants, apicomplexans, dinoflagellates, ciliates and diatoms for replication, transcription, splicing and translation

| PROCESS | Mammal | Plant | Alveolata | | | Diatom |
|------------------------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Subunits | | | Apico | Dinofl | Cilia | |
| | <i>Homsa</i> | <i>Arath</i> | <i>Plafa</i> | <i>Linpo</i> | <i>Tetth</i> | <i>Thaps</i> |
| <hr/> | | | | | | |
| DNA REPLICATION | | | | | | |
| DNA Polymerase | | | | | | |
| α complex | 4 | 4 | 4 | 4 | 4 | 4 |
| δ complex | 4 | 4 | 2 | 2 | 2 | 2 |
| ε complex | 4 | 2 | 2 | 2 | 2 | 3 |
| MCM complex | 6 | 6 | 6 | 6 | 6 | 6 |
| RPA | 3 | 2 | 1 | 1 | 1 | 2 |
| Clamp/loader | 4 | 4 | 4 | 4 | 4 | 4 |
| Other | | | | | | |
| Helicase | 1 | 1 | 0 | 1 | 1 | 0 |
| RNaseH1 | 3 | 3 | 1 | 0 | 3 | 1 |
| Fen1 | 1 | 1 | 1 | 1 | 1 | 1 |
| DNA ligase | 1 | 1 | 1 | 1 | 1 | 1 |
| <hr/> | | | | | | |
| TRANSCRIPTION | | | | | | |
| RNA Polymerase I, II and III | | | | | | |
| Core | 10 | 9 | 10 | 10 | 9 | 10 |
| Specific | 13 | 12 | 6 | 6 | 6 | 10 |
| Common | 5 | 5 | 4 | 5 | 4 | 4 |
| Basal Transcription Factors | | | | | | |
| TFIIA | 2 | 2 | 0 | 0 | 0 | 0 |
| TFIIB | 1 | 1 | 1 | 0 | 0 | 0 |
| TFIID | 15 | 10 | 1 | 1 | 3 | 4 |
| (TBP) | 1 | 1 | 1 | 0 | 1 | 1 |
| TFIIE | 2 | 2 | 0 | 0 | 0 | 1 |

| | | | | | | |
|-------------|----|----|---|---|---|---|
| TFIIF | 3 | 2 | 0 | 0 | 0 | 0 |
| TFIIH (NER) | 10 | 10 | 9 | 3 | 5 | 8 |

TRANSLATION

Ribosome

| | | | | | | |
|-------|----|----|----|----|----|----|
| B | 0 | 7 | 0 | 0 | 1 | 1 |
| B/A/E | 47 | 87 | 40 | 51 | 34 | 48 |
| A/E | 25 | 25 | 20 | 24 | 21 | 23 |
| E | 12 | 12 | 10 | 10 | 8 | 12 |

A.A-tRNA synthesis

| | | | | | | |
|---------|----|----|----|----|----|----|
| Enzymes | 23 | 23 | 22 | 22 | 22 | 22 |
|---------|----|----|----|----|----|----|

Basic translation factors

| | | | | | | |
|------------|----|----|----|----|----|----|
| Initiation | 42 | 57 | 29 | 29 | 33 | 37 |
| Elongation | 10 | 16 | 8 | 9 | 8 | 9 |
| Release | 11 | 11 | 5 | 3 | 7 | 4 |

SPLICING

Spliceosome

| | | | | | | |
|---------------|----|----|---|---|---|----|
| General | 9 | 8 | 8 | 9 | 9 | 7 |
| U1 | 8 | 7 | 5 | 5 | 5 | 4 |
| U2 | 12 | 10 | 7 | 8 | 9 | 10 |
| U4/U6 | 7 | 7 | 6 | 7 | 7 | 7 |
| U5 | 8 | 8 | 6 | 7 | 7 | 7 |
| U5/U4/U6 | 5 | 5 | 4 | 2 | 4 | 5 |
| Prp19 complex | 9 | 8 | 7 | 5 | 7 | 7 |
| Prp19 related | 9 | 8 | 8 | 7 | 5 | 8 |
| EJC/TREX | 6 | 5 | 4 | 3 | 4 | 5 |
| Common | 3 | 3 | 1 | 1 | 2 | 1 |

TRANSLATION RELATED

Ribosome biogenesis

| | | | | | | |
|------------------|----|----|----|----|----|----|
| 90S pre-ribosome | 18 | 18 | 6 | 12 | 15 | 14 |
| Nucleus | 14 | 12 | 9 | 10 | 12 | 12 |
| Nucleolus | 17 | 15 | 13 | 14 | 14 | 14 |

| | | | | | | | |
|---------------------------|----|------------|------------|------------|------------|------------|------------|
| Cytoplasm | | 7 | 6 | 5 | 6 | 6 | 6 |
| mRNA transport | | | | | | | |
| Nucleus | 11 | 9 | 9 | 7 | 7 | 8 | |
| Cytoplasm | | 6 | 6 | 2 | 3 | 2 | 3 |
| NPC | | 34 | 22 | 3 | 5 | 8 | 13 |
| SMNC | | 9 | 2 | 0 | 0 | 0 | 1 |
| eIFs | | 14 | 10 | 9 | 9 | 7 | 11 |
| EJC | | 16 | 12 | 6 | 9 | 6 | 6 |
| TREX | | 6 | 6 | 0 | 0 | 2 | 3 |
| mRNA surveillance pathway | | | | | | | |
| Nucleus | 34 | 26 | 9 | 15 | 10 | 17 | |
| Cytoplasm | | 15 | 12 | 8 | 10 | 7 | 9 |
| TOTAL | | 540 | 545 | 323 | 349 | 342 | 396 |

Table 3.ST3. Nuclear- and plastid-encoded reference sequences from GenBank used for comparison of synonymous and non-synonymous mutations.

| Gene Name | Accession | Length | Syn (dS) | Non-Syn (dN) |
|---------------------------------|-----------|--------|------------|--------------|
| p43 | AY423581 | 1429 | 14 (0.05) | 33 (0.04) |
| Phosphoribulokinase | AY772247 | 1461 | 64 (0.22) | 32 (0.03) |
| Histone like protein | AF482694 | 511 | 13 (0.17) | 18 (0.08) |
| Luciferase | AF085332 | 4000 | 25 (0.03) | 45 (0.02) |
| GAPDH (plastid isoform) | AF028560 | 1433 | 150 (0.47) | 35 (0.04) |
| Actin | AY423582 | 1407 | 59 (0.23) | 15 (0.02) |
| Rubisco | GONR15B | 1912 | 136 (0.36) | 33 (0.03) |
| Carbonic anhydrase | EU044834 | 1636 | 18 (0.06) | 23 (0.02) |
| Cyclin | AY618995 | 1825 | 6 (0.02) | 30 (0.03) |
| Cellulase | GQ258705 | 1425 | 40 (0.12) | 30 (0.03) |
| Glucose phosphate isomerase | DQ812892 | 1875 | 10 (0.03) | 18 (0.01) |
| Fructose 1,6 biphosphate | DQ508159 | 1235 | 12 (0.05) | 32 (0.04) |
| Sedoheptulose 1,7 biphosphate | DQ508153 | 1492 | 58 (0.18) | 44 (0.05) |
| Superoxide dismutase | AF289824 | 744 | 21 (0.13) | 5 (0.01) |
| Peridinin Chlorophyll a Protein | JO692699 | 1127 | 7 (0.025) | 11(0.013) |
| Luciferin Binding Protein | GONLBPA | 2217 | 37 (0.08) | 62 (0.04) |
| psaA | DQ264850 | 2506 | 72 (0.14) | 273(0.16) |
| psaB | DQ264852 | 2174 | 41 (0.09) | 146 (0.09) |
| psbA | DQ264844 | 1074 | 10 (0.04) | 32 (0.04) |
| psbB | DQ264845 | 1559 | 29 (0.08) | 103 (0.09) |
| psbC | DQ264846 | 1418 | 43 (0.17) | 112 (0.14) |
| psbD | DQ264847 | 1236 | 11 (0.04) | 55 (0.06) |
| atpA | DQ264853 | 1609 | 20 (0.06) | 69 (0.07) |
| atpB | DQ264857 | 771 | 17 (0.10) | 82 (0.14) |
| petB | DQ264849 | 842 | 21 (0.13) | 29 (0.06) |
| petD | DQ264848 | 545 | 13 (0.11) | 48 (0.13) |

3.6. Acknowledgements

We thank S. Kholmogorova and S. Benribague for helpful discussion, and Dr. Tao Jin (BGI) for review of the manuscript. Financial support of the National Science and Engineering Research Council (NSERC) of Canada to DM and BFL (Grant numbers 171382-03 and 194560) is gratefully acknowledged. SD-B and MB are studentship recipients from FCAR and NSERC respectively.

CHAPTER 4 – PUBLICATION # 3

Predicted Casein Kinase 2 sites in RNA binding proteins of *Lingulodinium* show daily variations in phosphorylation state

Sougata Roy¹ and David Morse¹ (Submitted)

¹ *Institut de Recherche en Biologie Végétale*

Département de Sciences Biologiques, Université de Montréal

4101 Sherbrooke est, Montréal, Québec, Canada H1X 2B2

4.1. Abstract

Many cellular processes in the dinoflagellate *Lingulodinium*, including bioluminescence, photosynthesis, cell division and nitrate metabolism, are controlled by a circadian (daily) clock. Since the activity of proteins involved in various metabolic pathways or in regulating gene expression can be affected by phosphorylation, we performed a comparative analysis of the phosphoproteome in early day (ZT2) and early night phase (ZT14) cells using phosphoproteins purified from these two different times. Long column LC-MS/MS identified over 10,000 peptides, of which 527 had at least one identified phosphosite and were derived from approximately 470 proteins. GO analysis of these proteins revealed RNA binding and translation as one of the major categories found. To identify phosphoproteins that might be substrates for kinases known to be important in eukaryotic circadian biology (CK1, CK2, AMPK, GSK3 β) we first categorized the kinases in the *Lingulodinium* transcriptome then assigned the different phosphosites to the different kinase classes. CK2 substrates were of particular interest as it is involved in all eukaryotic circadian systems known to date. Potential CK2 targets included several RNA binding proteins, one of which showed a 1000-fold difference in phosphopeptide abundance between the two different times. Apart from revealing a plethora of phosphoproteins involved in different metabolic processes, these analyses also provide a promising new approach to investigate the *Lingulodinium* circadian system, since regulation of RNA binding activity at different times could be used to regulate circadian-controlled gene expression in *Lingulodinium*.

Key Words: Circadian rhythms, Dinoflagellate, *Lingulodinium*, Phosphoproteomics, Post-Translational modification, Translational control

4.2. Introduction

Lingulodinium polyedrum (previously known as *Gonyaulax polyedra*) is a unicellular, photosynthetic dinoflagellate popularly known for an ability to produce red tides [433] and nightly bioluminescence [434]. Dinoflagellates, along with diatoms, are among the most important primary producers in the ocean [131], and those species that associate with the corals are vital for ocean ecology and biodiversity [432]. *Lingulodinium* has been principally studied as a model for understanding the biochemical foundations of the many physiological rhythms whose timing is regulated by the circadian clock [71]. At least in part, this regulation occurs through a control of gene expression, a facet of considerable interest in dinoflagellates not only from a circadian viewpoint but also from a standpoint of basic cell biology. Little is currently known about regulation of gene expression in dinoflagellates. Transcriptional control in particular is problematic, as dinoflagellate chromatin is organized in a cholesteric liquid crystalline structure [116] and contains no demonstrable histone proteins [184], although in *Lingulodinium* all the core histone genes are transcribed and their sequences are well conserved [189]. Furthermore, the recently described *Lingulodinium* transcriptome shows a severe under-representation of all known DNA binding protein (DBP) classes. Lastly, genes known to be regulated by the circadian clock appear to date to be all regulated at a translational rather than a transcriptional level [165].

In cases where post-transcriptional regulation appears prevalent, proteomics is emerging as a useful tool to address the end result of altered gene expression, and can provide valuable insight into functioning of metabolic pathways important for the survival and fitness of an organism [463]. While levels of proteins often play the major role, post-translational modifications (PTM) that modulate the structure and function of proteins [365] can also be involved. Phosphorylation is one of the most widely used and well-characterized PTM and can affect protein folding, enzyme activity, interactions between proteins, degradation rates and sub-cellular localization [367], which in turn can be used to regulate a wide variety of cellular activities such as intercellular communication, growth, proliferation, differentiation and apoptosis [369]. With the advent of sophisticated mass spectrometric instruments [464, 465] and improved phosphopeptide enrichment techniques [466], large-

scale and comprehensive analysis of phosphoproteomes has now become feasible. Indeed, phosphopeptides can not only be identified, but their levels quantified, thus allowing a comparison of peptide levels during different conditions [467]. A screening of phosphopeptide intensities at different times *in vivo* could thus identify any proteins that respond to different times of day by a differential phosphorylation and that might be involved in the regulation of diverse physiological rhythms by changes in activity rather than in protein amount. Proteomics and phosphoproteomics have remained largely unexplored in dinoflagellates, in part because of difficulties in phosphoprotein enrichment techniques and in part due to the paucity of sequence databanks. However, since dinoflagellates are not yet amenable to genetic transformation or mutational studies, proteomic approaches constitute particularly promising research avenues.

Phosphorylation is also an integral part of circadian time keeping in animal, plant, fungal and cyanobacterial models [81, 86, 371, 372]. Circadian biology involves study of circadian rhythms, the roughly 24h rhythms in cellular physiology, behavior or metabolism that allow cells to perform specific tasks better at specific times of day, as well as circadian clocks, the endogenous timers that provide the signals required to orchestrate the different rhythms [63]. The biochemical basis for circadian rhythms is not well understood, and indeed, some of the best-studied examples are found in *Lingulodinium*. In contrast, the central core of the clock mechanism has been intensively studied in model systems and is completely unknown in *Lingulodinium*. The clock involves a coupled transcription/translation negative feedback loop where specific transcription factors activate expression of factors that inhibit their own transcription [468]. However, translational regulation mediated by RNA binding proteins (RBPs) has also been shown to affect the clock mechanism [469, 470], and the phosphorylation state of core clock proteins at different times also appears to be of critical importance to the proper functioning of the clock [86, 371].

The extent to which phosphorylation is used by the clock to regulate biological rhythms as opposed to the clock's time-keeping mechanism itself is still unclear. In *Arabidopsis*, transcription of several kinases and phosphatases are under circadian regulation

and their profile changes according to conditions [373, 374], and thus some of these may regulate rhythms. On the other hand, kinases are essential components of the clock mechanism [471], and among kinases, the Casein Kinase 2 (CK2) is particularly important as a ubiquitous and evolutionary conserved clock component in eukaryotes [89, 91, 471, 472]. Involvement of a CK2 in the circadian systems of other species remains to be determined, although in *Lingulodinium*, serine/threonine kinase inhibitors are known to affect the timing of the bioluminescence rhythm [93, 94] suggesting phosphorylation will play a role in this clock as well.

The present study was aimed at the identification of *Lingulodinium* proteins whose phosphorylation state varied over time, with a particular emphasis on potential CK2 targets. In order to maximize the number of differentially phosphorylated peptides, this analysis used cells taken during a light/dark cycle and will thus include those proteins differentially phosphorylated as a direct result of the light/dark cycle or by the circadian clock, either as clock components or for mediating clock control over the biological rhythms. With respect to the latter, differentially phosphorylated RBPs might be involved in regulating the synthesis rates of proteins whose translation has been previously demonstrated to be clock controlled [347]. We chose ZT2 (early day) and ZT14 (early night) for comparative phosphoproteome analysis as PCP (peridinin-chlorophyll a-binding protein), the major light harvesting protein is translated at ZT2 and not at ZT14, while LBP (luciferin binding protein), a substrate binding protein in the bioluminescence reaction is synthesized at ZT14 and not at ZT2. We found that CK2 appears to be an important kinase in *Lingulodinium* with respect to the number of its predicted target sites and that many RBPs contain these sites.

4.3. Materials and methods

4.5.1. Cell Culture

Lingulodinium polyedrum cultures (formerly *Gonyaulax polyedra*; strain CCMP1936) were obtained from the Provasoli-Guillard Culture Center for Marine Phytoplankton (Boothbay Harbor, Maine. Cells were grown in f/2 seawater medium lacking Si [399, 473] at constant temperature ($19 \pm 1^\circ\text{C}$) under 12-h light/12-h dark cycles using cool white fluorescent light at an intensity of $50 \mu\text{mol photons m}^{-2}\cdot\text{s}^{-1}$. Under this light regimen, ZT 0 corresponds to the beginning of the light period and ZT 12 the beginning of the dark period. Cultures were grown to a cell density of roughly 10^4 cells/mL then harvested by filtration on a Whatman 541 paper supported by a Buchner funnel. Cells were washed twice with 250 ml phosphate free seawater, and then frozen in liquid nitrogen and either used immediately or stored at -80°C until use.

4.5.2. Phosphoprotein purification and gel electrophoresis

Phosphoproteins were purified from *Lingulodinium* crude protein extracts prepared at both ZT2 and ZT14 using a PhosphoProtein purification kit (Qiagen) following manufacturer's instruction. Briefly, frozen cells were crushed in liquid nitrogen in a mortar and pestle, immediately put in the supplied lysis buffer (Qiagen), and incubated on ice for 30 min with occasional mixing. Insoluble cell debris was removed by two sequential centrifugations at 15,000 g for 10 and 5 minutes respectively, each at 4°C , and the supernatant retained. Protein concentrations were measured using Bradford assay (BioRad) and using a VersaMax (Molecular Devices) plate reader. Total protein (2.5 mg) was diluted to a final concentration of 0.1 mg/mL in 25 mL of lysis buffer (supplemented with benzonase and protease inhibitors) and passed through the affinity column at room temperature to capture the phosphoproteins. The flow-through was collected and kept for further analysis. Remaining unbound proteins were removed by washing with the lysis/wash buffer and then phosphoproteins were eluted with the supplied elution buffer (Qiagen). The eluted

phosphoproteins were desalted and precipitated with 4 volumes of prechilled (-20°C) acetone for 2 hours at -20 °C. To assess the efficiency of the purification, the initial total protein extract and the flow-through were desalted and dissolved in SDS buffer (2% SDS, 0.7mM 2-mercaptoethanol, 62.5mM Tris-HCl, pH 6.8, 10% glycerol) while the dried pellet of the specific eluent from the column was dissolved directly in SDS buffer. All samples were heated at 95°C for 3 min and electrophoresed on 12.5%PAGE containing SDS. The gel was sequentially stained with ProQ diamond (Invitrogen) for phosphoproteins and Sypro Ruby (BioRad) for total proteins respectively, following the manufacturer's protocols. The fluorescence emission from ProQ Diamond and Sypro Ruby stained gels were captured using a Typhoon PhosphorImager (GE Healthcare).

4.5.3. Mass Spectrometric analysis

The acetone precipitated enriched phosphoprotein samples from *Lingulodinium* were used directly for trypsin digestion and LC-MS/MS analysis at the proteomic facility of l'Institut de recherche en immunologie et en cancérologie (IRIC, Université de Montréal, Canada). All samples were prepared for digestion by resuspension in 50 µl 50 mM ammonium bicarbonate and TCEP (tris (2-carboxyethyl) phosphine) was added to reach a final concentration of 5mM. Samples were incubated at 37 °C for 30 min, then 30 µl 55 mM chloroacetamide was added and the samples incubated for a further 30 min at 37 °C. Samples were digested overnight at 37 °C in the presence of 1µg of trypsin, then dried in a Speed-Vac and resolubilized in 50 µl of 5% acetonitrile/ 0.2% formic acid. 20 µl of each samples were injected on a C18 precolumn (0.3 mm i.d. x 5 mm) and peptides were separated on a C18 analytical column (150 µm i.d. x 100 mm) using an Eksigent nanoLC-2D system. A 56-min gradient from (A/B) 10–60% (A: formic acid 0.2 %, B: acetonitrile/0.2% formic acid) was used to elute peptides with a flow rate set at 600 nanoliter/min. The LC system was coupled to a LTQ-Orbitrap Velos mass spectrometer (Thermo Fisher). Each full MS spectrum was followed by 12 MS/MS spectra (thirteen scan events), and the 12 most abundant multiply charged ions selected for MS/MS sequencing. Tandem MS experiments were performed using collision-induced dissociation in the linear ion trap. Data were processed with the Mascot 2.2 (Matrix Science) search engine using a sequence assembly derived from a *L.*

polyedrum transcriptome [165] assembled from ~300 million 76 bp Illumina paired-end reads using either Velvet and Oases (JO692619–JO767447) or Trinity (this Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GABP000000000. The version described in this paper is the first version, GABP010000000). The variable modifications included were deamination [474], carbamidomethyl (C), oxidation (M) and phosphorylation (STY). Precursor and fragment tolerances were 10 ppm and 0.5 Da, respectively.

The comparison of peptide abundances between samples was performed using raw data files (.raw) from the Xcalibur software, which were first converted into peptide map files representing all ions according to their corresponding m/z values, retention times, intensities, and charge states. Intensity values above a threshold of 10,000 counts were considered for further analysis. Peptide abundances were assessed using the peak top intensity values. Clustering of peptide maps across different sample sets were performed on peptides associated to a Mascot entry using hierarchical clustering with tolerances of 15 ppm and 1min for peptide mass and retention time, respectively. Retention time of the initial peptide cluster list was normalized using a dynamic and nonlinear correction to confine the retention time distribution to less than 0.1 min (<0.3% RSD) on average.

The variation of intensities between samples was used to compute the fold change of a protein. First, a number between 0 and 1 that described the amount of representation of the protein within each condition was assigned. Then an in-house software (ProteoProfile; <http://www.thibault.irc.ca/proteoprofile/files/TechnicalGuide.pdf>), which assigns weights to the peptides composing the protein, was used to calculate the relative intensities for each protein. The weight of each peptide represents its potential to describe correctly the protein. Each peptide starts with a weight in proportion to its own intensity level (Log 10 of the average intensity of the peptide divided by 10). Based on the Weiszfeld's iteratively re-weighted least squares algorithm, this weight is multiplied by the closeness of the peptide to the protein's fold change, through a series of iterations.

4.5.4. Bioinformatic Analysis

The LC-MS/MS analysis returned a list of peptides along with their relative intensities at ZT2 and ZT14. Some peptides appear several times in the list, therefore intensities of the peptides with identical sequences were summed to yield final peptide intensity. For some analyses, all peptides identified by comparison to the *Lingulodinium* transcriptome were used as all are potentially derived from phosphorylated proteins. However, for most analyses only peptides with an identified phosphosite(s) were used.

Sequence annotation and mapping to Gene Ontology (GO) [441] and Interpro domains [475] were performed using the web based tool Blast2GO [443]. Blast2Go default parameters were used to perform blastX and annotate the total transcriptome and the enriched protein fraction into different GO categories. For annotating the ZT2 and ZT14 hyperphosphorylated proteins, the Interpro domain information was used to infer a function for some sequences where a GO category for the protein itself was not available. Information from these two different sources was merged and verified manually to obtain the largest number possible of identified proteins.

To determine the repertoire of *Lingulodinium* kinases, protein sequences corresponding to all available kinase classes (AGC, CAMK, CMGC, CK-1, Atypical, STE, TKL, TK and Others) were downloaded from KinBase (<http://kinase.com/kinbase/>). Each group of kinase sequences were stored separately as a local database in the Geneious software program [476] and were screened with the *Lingulodinium* transcriptome using BLASTx with a cut-off of $< e^{-25}$. This search returned different sets of *Lingulodinium* sequences belonging to the different kinase classes. We realized that this list might contain several kinases with multiple representations in different kinase groups. Hence each group of *Lingulodinium* kinase sequences were then compared back to the 'All kinome proteins' database using the BLASTx program available in the kinase.com site, which provides a hierarchical classification of each kinase. All other settings were kept at default values. The *Lingulodinium* kinases were classified by the best BLAST hit, and sequences

found in more than one kinase class were only included in one category (with the best E-value). The *Lingulodinium* kinase datasets were then compared to the human [477], sea urchin [478] and *Tetrahymena* [479] kinase classes downloaded from kinBase.

A Group-based Prediction System (GPS version 2.1.2) software (<http://gps.biocuckoo.org>) was used for computational prediction of protein kinases capable of phosphorylating the different phosphosites [480] identified by MS sequencing. Only kinases present in *Lingulodinium* were selected from the list available within the program and the threshold was maintained at high in order to retrieve only the most likely kinase candidates. To avoid redundancy, only a single kinase was retained for each phosphosite on the basis of best score and cut-off values. The prediction program also predicted kinases for many non-phosphorylated S and T present in the peptide, and these were manually removed so that only the experimentally determined phosphosites would be analysed further.

4.4. Results

4.3.1 Phosphoprotein purification yields more peptides than phosphopeptide enrichment

A previous attempt to analyze the *Lingulodinium* phosphoproteome using TiO₂ phosphopeptide enrichment yielded a large number of acidic peptides of which only a tenth were bona fide phosphopeptides (Table 4.1.) [376]. To attempt to reduce the contamination by acidic peptides, we enriched for phosphoproteins instead of phosphopeptides. A commercial system (QIAGEN), used in many other systems with satisfactory results [288, 481-483], was first assessed by one-dimensional SDS-PAGE (Fig. 4.1.). We find almost all Pro Q diamond (phosphoprotein) staining in the specific eluent (lane 3) and little in the column flow-through (lane 2). In contrast, Sypro Ruby (general protein) staining of the flow-through and the specific eluent shows similar amounts of protein but markedly different protein profiles. We conclude that this protocol selectively captures the bulk of *Lingulodinium* phosphoproteins from our crude extracts and recovers them in an eluted fraction with a decreased protein complexity.

The analysis of the enriched proteins was performed as described [376] except for the use of a long column (for better separation and resolution of peptides) instead of a regular C18LC step and the use of a sequence database combining a previously described Velvet assembly [165] with a newer Trinity assembly (the Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GABP000000000. The version described in this paper is the first version, GABP010000000) from the same reads. Our new protocol resulted in a roughly ten-fold increase in both total number of peptides and number of phosphorylated peptides identified when compared to the previous phosphopeptide enrichment (Table 4.1.) [376]. The total number of proteins from which the different identified peptides originated (3007) suggests that the enriched phosphoproteins constitute roughly 4% of the *Lingulodinium* sequences in our assembly (74,655). This is much less than the phosphoprotein content of higher eukaryotes, estimated to be close to 30% [484], and this may be in part due to the large number of short sequences in our

assembly which tends to exaggerate the estimated number of proteins to 74,655. Of the total peptides identified in the phosphoprotein fraction, only 5% were found to contain one or more phosphate moieties by MS sequencing, a value similar to the 360 phosphopeptides obtained from 328 phosphoproteins in *Chlamydomonas* [485]. The ratio of different phosphorylated amino acids in *Lingulodinium* is similar to that in *Arabidopsis* and human, although we find 14% more pThr, 11% less pSer and slightly less pTyr (2%) [486-489].

4.3.2. The phosphoproteome fraction is enriched in proteins involved in translation and RNA binding

The 11,188 peptides identified from the two times combined were derived from 3,007 proteins. Roughly one-quarter of these proteins could be annotated by function using Blast2GO, a similar proportion to those annotated using the whole transcriptome [165]. To assess if our enriched phosphoprotein fraction contained over-represented or under-represented classes, GO category analysis for molecular function, biological process and cell component of the enriched fraction was compared to that of the whole *Lingulodinium* transcriptome (Fig. 4.2.). Interestingly, the phosphoproteome fraction is enriched in RNA and nucleotide binding proteins (Fig. 4.2.A), in proteins involved in translation or gene expression (Fig. 4.2.B) and for proteins with a cytoplasmic location (Fig. 4.2.C). In contrast, DNA binding proteins, as shown by both transcription (Fig. 4.2.B) and nuclear localization (Fig. 4.2.C) categories, are severely under-represented. These results thus differ from those in a large-scale comparative phosphoproteome using Rice and *Arabidopsis*, which clearly showed enrichment in DNA binding, transcription and nuclear localization [490]. Our results also show no enrichment for signal transduction, response to stress and stimuli, protein modification or plasma membrane protein categories, again dissimilar from what is observed for plant phosphoproteomes [490]. We do observe kinases as more highly represented category in the *Lingulodinium* phosphoproteome fraction (Fig. 4.2.A), but this is not as marked a difference as is found for plant phosphoproteomes [490].

4.3.3. Phosphopeptide intensity comparisons between ZT2 and ZT14 reveal many RNA binding proteins

The phosphopeptide intensities reflect the amount of a specific phosphopeptide and will be influenced by both the total amount of protein and the degree of phosphorylation. However, to date, no major changes in protein levels have been observed between early day and early night cultures [347]. Thus, to a first approximation, we consider this phosphopeptide abundance a reasonable indicator of differential phosphorylation. To then identify proteins that could be differentially phosphorylated at the two times, we plotted the signal intensity of 527 phosphopeptides (corresponding to approximately 470 different phosphoproteins) from the ZT2 sample against those for ZT14 (Fig. 4.3.). Note that in this analysis, only peptides with *bona fide* phosphate signatures were used, reasoning that even though the majority of the peptide sequences in the phosphoprotein fraction are likely derived from phosphoproteins, changes in peptide intensities due to other post-translational modifications cannot be ruled out. We classified peptides with more than 2-fold variation in signal intensity as hyperphosphorylated at one of the two times, with the remaining peptides considered as unchanged. We found 170 (32%) and 130 (25%) peptides to be hyperphosphorylated at ZT2 and ZT14, respectively. The average change in phosphopeptide levels is difficult to estimate as intensity values below threshold (10,000) are not recorded. However, if all threshold values are arbitrarily set to 10,000, the average phosphopeptide intensity is roughly five times greater at ZT2 than at ZT14. Fig. 4.3. also suggests increased levels of phosphorylation at ZT2, as more peptides are found directly on the Y-axis and intensities are spread out over a larger range than for the x-axis.

After identifying all the proteins containing the 527 phosphopeptides (Supplementary Table 4.S2.), we next grouped and thoroughly characterized the proteins whose peptides were hyperphosphorylated at one of the two times (Fig. 4.4.). Interestingly, 8% of the identifiable hyperphosphorylated proteins were classified as RNA-binding proteins [347] a proportion even higher than that found in the ensemble of enriched phosphoproteins (5.3%). The peptide intensity of RNA binding proteins can vary by up to 1000-fold between the two

times studied. We note that an additional 10 proteins involved in various aspects of global translation were also differentially phosphorylated between these two times. Furthermore, four zinc finger domain-containing proteins, which are often transcription factors or involved in nucleic acid binding, are hyperphosphorylated (30 to 1000-fold) at ZT2 (Table 4.2.). A total of 23% of the differentially regulated phosphoproteins were involved in general metabolism, a quarter of which were related to amino acid metabolism pathways.

4.3.4. Orthologs of kinases involved in circadian regulation in other eukaryotes may regulate translation in *Lingulodinium*

The total number of kinases as well as their distribution among known kinase classes has not been previously assessed in dinoflagellates, so we first performed this analysis for the entire transcriptome (Supplementary Table 4.S1.). From a total of 74,655 sequences, 611 (0.7%) code for kinases. Among the kinases, the cyclic AMP-dependent kinase (CAMK) class and the calcium-dependent protein kinase (CDPK) sub-group in particular were found to be over-represented. In contrast, the tyrosine kinase group (TK) is severely under-represented, as is also the case in other plants and algae [491].

Having classified the *Lingulodinium* kinases, we next used the GPS software [480] to predict which kinases were likely to have phosphorylated the different phosphosite motifs (Fig. 4.5.A). The TKL, CMGC and ‘other’ kinase groups appear responsible for most (roughly 75%) of the phosphorylation events. The CK2 subclass in particular (found in the “other” kinase group) is predicted to phosphorylate 14% of the total phosphosites, making it the largest single group in *Lingulodinium*. CK2 kinases are predicted to phosphorylate more peptides at ZT2, while kinases in the CAMK and atypical kinase groups appear to be more active at ZT14. However, there is no correlation between the number of kinases and the number of sites phosphorylated per kinase class among the peptides found here (Fig. 4.5.B).

We were particularly interested in Casein Kinase 1 (CK1 δ/ϵ), Casein Kinase 2 (CK2 α and β), Glycogen synthase kinase 3 β /shaggy-like kinase (GSK3 β) and 5' adenosine monophosphate-activated protein kinase (AMPK), as these have all been shown to regulate core clock proteins in different eukaryotic model systems [83, 492-494]. Among these, CK2

has been implicated as a clock component in *Drosophila* [89], *Neurospora* [92], plants [472] and mammals [90], so we first looked for the protein targets of this kinase in *Lingulodinium*. Almost one-third of the predicted CK2 targets are either DNA/RNA binding or involved in nucleic acid structure and processing, a proportion similar to the mammalian CK2 substrates that are involved in gene expression and protein synthesis [495]. Some are proteins with multiple functions, such as glyceraldehyde-3-phosphate dehydrogenase (GAPDH), which not only plays an important role in glycolysis but is also known to regulate turnover and translation of specific mRNAs [496, 497]. Other targets are more specifically related to control of gene expression, with five of the 32 target substrates identified as RBPs. One RBP in particular (accession number J0719969) is much more heavily phosphorylated (1000-fold) at ZT2 than at ZT14. This protein was also observed to be more phosphorylated at LD 6 than at LD18 during a previous study [376].

4.5. Discussion

The development of efficient phosphoprotein/ phosphopeptide enrichment protocols has allowed a detailed examination of the *in vivo* dynamics of protein phosphorylation in plant systems [483, 498]. However, this technique has not been extended to dinoflagellates, organisms that may greatly benefit from this type of analysis because of a reduced dependence on transcriptional regulation. Here we present a comparative analysis of a phosphoprotein enrichment protocol using extracts from the dinoflagellate *Lingulodinium* prepared at two different times, ZT2 (early day) and ZT14 (early night). This method provides a significant improvement over a previously tested TiO₂ phosphopeptide enrichment [376], as those phosphopeptides were heavily contaminated by acidic peptides which were not necessarily derived from phosphorylated proteins. In combination with the use of the long run LC, the phosphoprotein enrichment protocol identified roughly 25 times more peptides (corresponding to ten times more proteins identified) than the previous method (Table 4.1.). This greater proportional increase in the number of peptides after phosphoprotein enrichment shows that each protein produces more identifiable peptides, and supports the contention that the peptides are in the enriched fraction because they are part of a phosphorylated precursor and not because they share a particular chemical property (such as acidity). The fact that only a low proportion of the peptides recovered (5%) have a confirmed phosphosite may be due to a still meagre database support for our species. Indeed, our recently published transcriptome [165] is the most complete dataset available, yet the average transcript size in the assembly is much smaller than that of the average mRNA size. However, it is also likely that the many non-phosphorylated peptides derived from the phosphorylated protein purified will also contribute to a low percentage of phosphopeptides. Since it is not possible to determine if all peptides recovered do indeed belong to phosphoproteins, and to avoid counting the same phosphoproteins several times, we chose to perform most of our analyses on those peptides with an identified phosphosite.

Some of the differentially phosphorylated proteins may be involved in mediating control over general metabolism, as the largest number of phosphorylated targets are metabolic enzymes, including some related to amino acid metabolism, are more highly

phosphorylated during the early day. Interestingly, a soluble starch synthase isoform, a key enzyme in starch synthesis, is hyperphosphorylated at night (690-fold) while a protein related to glycogen phosphorylase, required for breakdown of stored glycogen, is more highly phosphorylated at day (600-fold). In many plants and green algae, starch synthesis occurs during the day [499] while glycogen/starch breakdown occurs at night in order to compensate for the absence of light energy [500]. Inhibition of the activities of starch synthase and glycogen phosphorylase like enzymes by phosphorylation might be a biologically relevant adaptation to light/dark cycles.

Surprisingly, few of the differentially enriched phosphopeptides were found to correspond to kinases and phosphatases when compared to the total number of these proteins in the transcriptome. We had originally thought that regulation of these classes might be important as a series of previous studies had emphasized the importance of both serine/threonine kinases and phosphatases in *Lingulodinium* rhythms [93-96]. These studies used inhibitors to affect canonical clock properties, thus suggesting a role in the clock mechanism itself. The results shown here allow a fine tuning of this idea, as they suggest it is not the kinases/phosphatases themselves but rather their targets that are more likely to be involved in regulating clock properties.

The class of differentially phosphorylated proteins of greatest interest are those potentially involved in binding RNA, as to date, all reports of gene regulation in *Lingulodinium* show a predominant dependence on translational control [341] and RNA binding proteins (RBPs) are one of the important executors of this mechanism [501]. RBPs generally affect protein synthesis by binding to untranslated regions [424] in mRNAs [502], and they can be regulated by phosphorylation events which either activate or inhibit their activity directly, or which cause effects on their stability or localization [503]. We found 17 different RBPs that were differentially phosphorylated at ZT2 and ZT14 (Table 4.2.), of which most (70%) were proteins containing RNA Recognition motif (RRM) domains. RRM domains are conserved motifs known to bind single stranded RNA [504-506] and several RRM domain-containing proteins are known to participate in translational regulation [507-510]. The next most abundant group were the K-homology (KH) domain containing proteins,

which also function in RNA binding and recognition [511] and in yeast, the translational repressor activity of the KH domain protein Khd1 was reduced when phosphorylated by CK1 [512]. Pumilio (Puf) domain proteins are well known for their role in translational regulation [513] through sequence specific binding of mRNAs [514, 515]. In yeast, phosphorylation of PUF6 by CK2 changes its binding affinity for ASH1 mRNA thereby relieving the translational repression [516]. We found only a single Pentatricopeptide (PPR) protein that was differentially phosphorylated (22-fold greater at ZT 2). PPR proteins also bind RNA but appear to regulate gene expression function in organelles [517]. Unfortunately, due to the incomplete 5' end of the PPR encoding transcript sequence in our transcriptome, we were unable to confirm the localization of the encoded protein. Lastly, we have found two cold shock domain proteins (CSPs), one hyperphosphorylated at each of the two times. CSPs are also known to bind RNA in bacteria [518] and plants [519] and have been shown to regulate translation in prokaryotes [520] as well as in eukaryotes [521-523].

We also report here for the first time the repertoire of kinases in *Lingulodinium* (Supplementary Table 4.ST1.) and have compared the kinase catalogue with those of other eukaryotes (Supplementary Figure 4.S1.). The over-representation of CDPK (a sub-group representing almost 3/4th of the CAMK class) in *Lingulodinium* agrees well with the earlier reports describing the important role of calcium signalling in different cellular processes of dinoflagellates [524]. For example, in the absence of mechanical stimulation calcium signalling can induce bioluminescence [525], suggesting an involvement in the nightly bioluminescence process. However, we also note that the elevated number of members in this group does not correspond to the paucity of phosphopeptides classified as being phosphorylated by this group of kinases (Fig. 4.5.B). A number of possibilities may provide an explanation for this. First, the kinases might be active at times different from those analyzed in this study. Also, it may be that some of the kinases in the group are unique to the dinoflagellates and thus the GPS software used for kinase prediction would not have been optimized for these. Lastly, although we selected the best-predicted kinase for each phosphosite, a given site has many candidate kinases and the true kinase may actually be classed lower in the predictions.

Among the GPS-predicted kinase sites, CK2 (classified in the ‘other’ category) sites were prevalent, phosphorylating almost three-quarters of the sites in this group. The majority of the predicted CK2 targets belonged to either unknown proteins or to a variety of different metabolic pathways, consistent with the pleiotropic effects of CK2 observed in a study with large-scale substrate analysis from all available source organisms [495]. However, CK2 is evolutionarily conserved [526] and is a component of the circadian clock that phosphorylates and modulates the stability, activity and/or localization of core clock proteins [82, 90]. CK2 is thus a promising candidate that could be involved in modulating clock proteins in *Lingulodinium*, and is of considerable interest since no circadian core proteins have yet been identified in the dinoflagellates. It is interesting that a substantial percent (15%) of the differentially phosphorylated CK2 substrates were classified as RBPs. One RBP in particular (accession number J0719969) is more heavily phosphorylated at ZT2 by three orders of magnitude (Fig. 4.6.), and this protein will come under intense scrutiny in further studies to assess a possible role in the dinoflagellate circadian timing mechanism. As a caveat, however, we note that this experiment cannot distinguish between an effect of the biological clock and a direct response to a change in the light intensity. It will thus be of interest in future studies to assess the differences between the two times under either constant light or constant dark conditions to determine if the changes are due to the circadian clock.

The experiments described here provide a marked improvement in phosphoprotein analysis compared to our previous trials and contribute valuable insights into the dynamics of *Lingulodinium* phosphoproteins. Earlier studies have underscored the importance of kinases for generation of rhythms in *Lingulodinium*, but until recently, no databases were available to analyse the repertoire of kinase targets. The recent development of the *Lingulodinium* transcriptome database [165] has thus opened the way to characterization of the phosphoproteome. We note an extensive enrichment of RNA as opposed to DNA binding proteins, and in particular, have found several RBPs among the predicted CK2 targets. Since CK2 is a clock protein conserved throughout the eukaryotic kingdom, these RBPs may be candidates for involvement in the core clock mechanism of dinoflagellates, similar to that of Chlamy1 in *Chlamydomonas* [469] and cold-inducible RNA-binding protein in mammals [470]. While the experiments shown here reflect only two times of the daily cycle, we have

now demonstrated an important proof of principle for expanding the phosphoproteomic analysis to the complete circadian cycle.

Table 4.1. Comparison of phosphopeptide and phosphoprotein enrichment protocols

| | Phosphopeptide ¹ Enrichment | Phosphoprotein Enrichment |
|---------------------------|---|------------------------------|
| Total enriched peptides | 422 | 11,188 |
| Total enriched proteins | 293 | 3,007 |
| Phosphopeptides | 54 | 527 |
| Phosphoproteins | 45 | 470 |
| Phosphosites (%S; %T; %Y) | 61 (78.3; 20;1.7) | 690 (74;24;2) |

¹ Phosphopeptide enrichment data from [376].

Table 4.2. Number of hyperphosphorylated RNA-binding proteins at either ZT2 or ZT14

| RBP Type | Number of proteins | |
|---------------------|--------------------|------|
| | ZT2 | ZT14 |
| Zinc finger | 4 | 0 |
| RRM/KH/OBfold/CSP | 6 | 8 |
| Pumilio/RAP/PPR | 3 | 0 |
| General translation | 5 | 5 |

Figure 4.1. Efficient enrichment of *Lingulodinium* phosphoproteins by affinity chromatography

The crude extract (lane 1), flow-through (lane 2) and the specific eluate (lane 3) was electrophoresed using 12.5% PAGE containing SDS and the gel then stained sequentially with ProQ diamond and Sypro Ruby stains. The protein quantity in the flow-through and the eluted fraction lanes is the same but is 10-fold less than in the crude extract lane. The arrows indicate the position of two molecular markers. Lane 1 contains approximately 50 μg protein, while lanes 2 and 3 contain approximately 10 μg of protein.

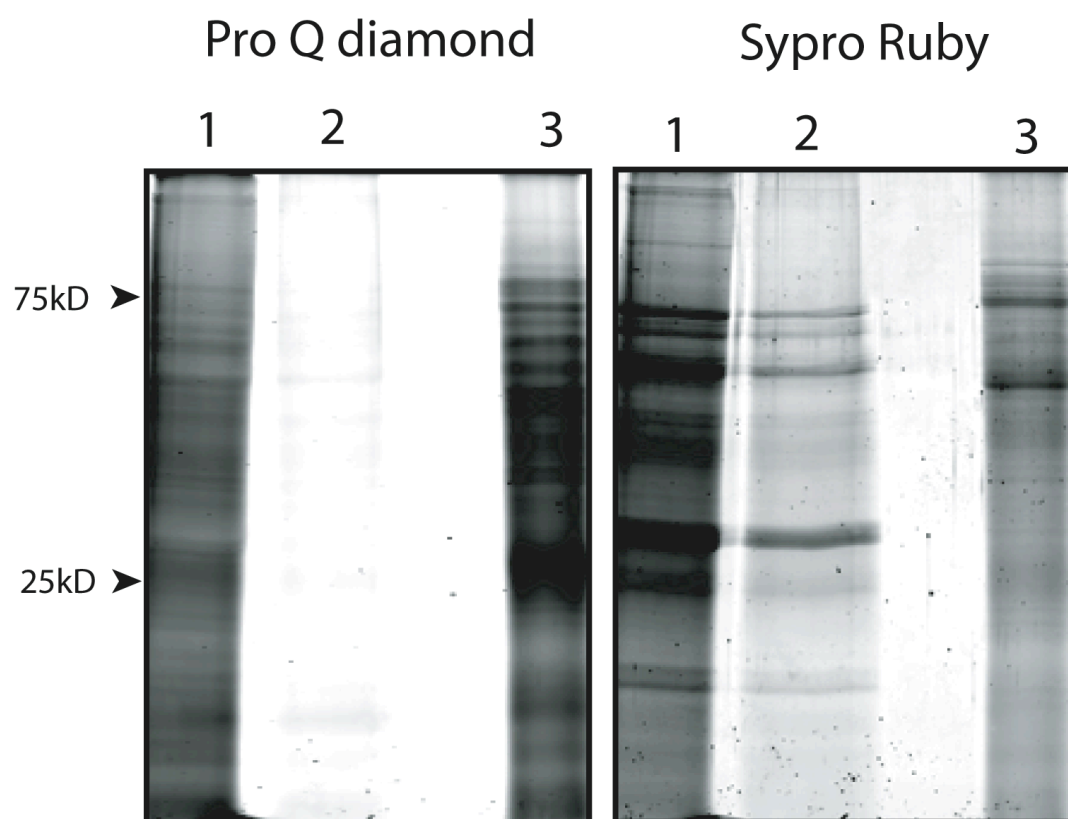


Figure 4.2. RNA rather than DNA related processes are highly represented in the total enriched phosphoprotein pool

The enriched phosphoproteins were classified into GO categories by molecular function (A), Biological Process (B) and Cell component (C) and compared to an analysis performed using the total transcriptome. Only those GO categories that were most informative in comparison are shown. Transport in (A) refers to Substrate-specific transmembrane transport; ER in (C) is Endoplasmic reticulum.

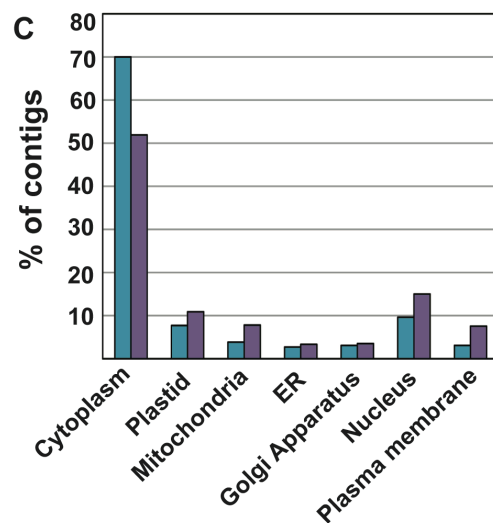
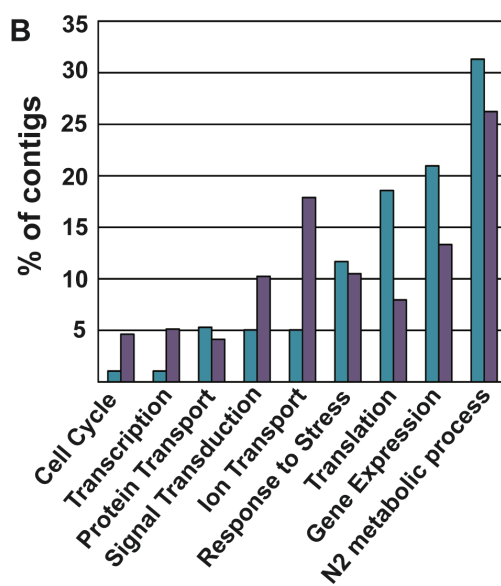
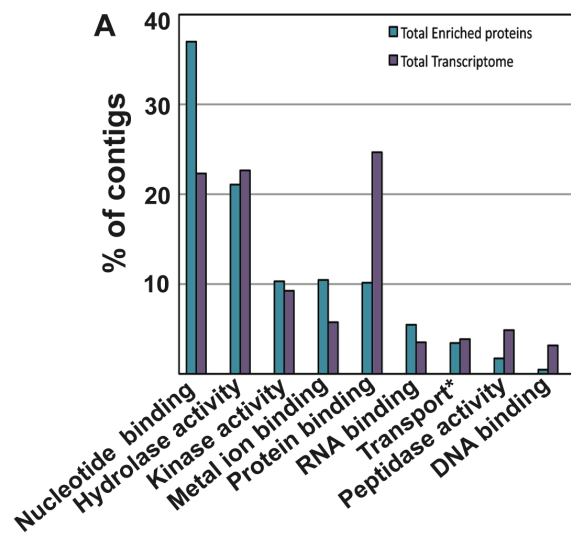


Figure 4.3. Phosphopeptide Intensity at ZT2 is much pronounced than at ZT14

The peptide intensities for each of 527 peptides at each of the two times are plotted against one another, with red and green dots denoting peptides that are hyperphosphorylated at ZT2 and ZT14 respectively. Peptides whose intensities change by less than 2 fold are shown in blue. Points lying on either axis (i.e. 10,000) are at or below threshold levels and thus actual fold changes cannot be ascertained.

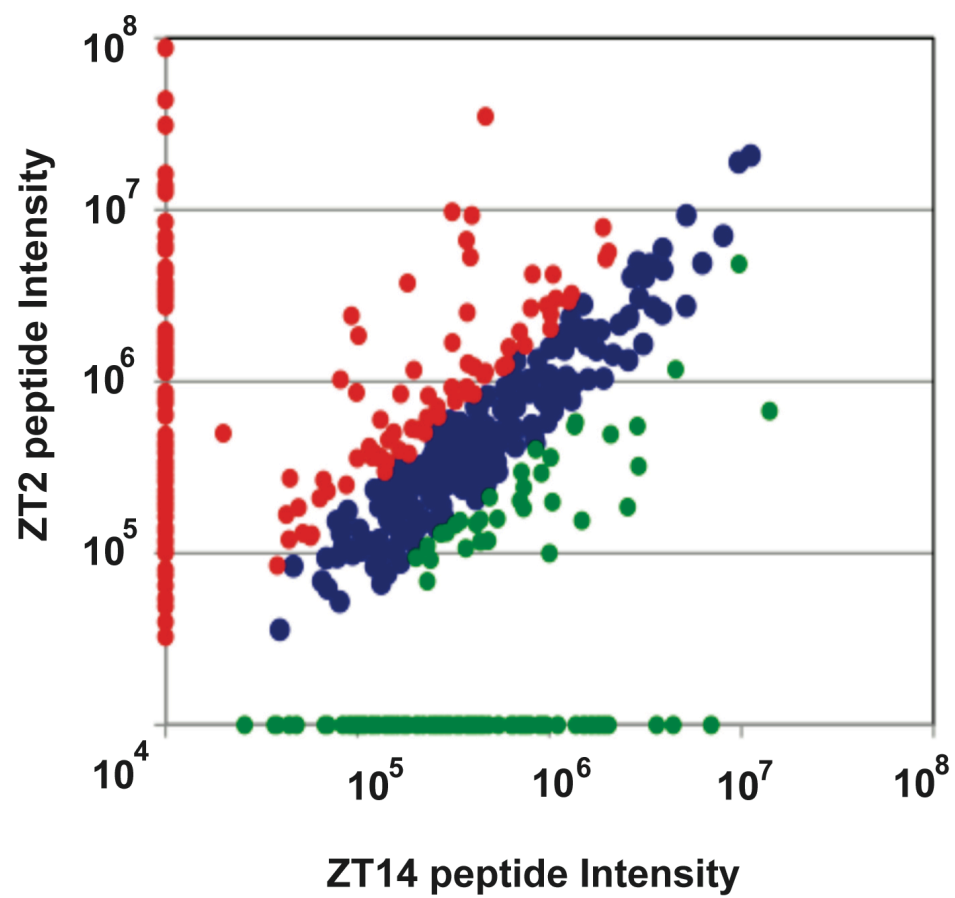


Figure 4.4. Many RBP are differentially phosphorylated at ZT2 and ZT14

All proteins containing hyperphosphorylated peptides at either of the two times were grouped (as described in Methods section). To emphasize the importance of RBPs, the ‘DNA and RNA related processes’ was replaced by ‘DNA/RNA binding’, which excludes translation factors or activators, proteins related to splicing, ribosome, transcription factors or tRNA activation. Red and green bars correspond to the peptides with greater intensities at ZT2 and ZT14, respectively.

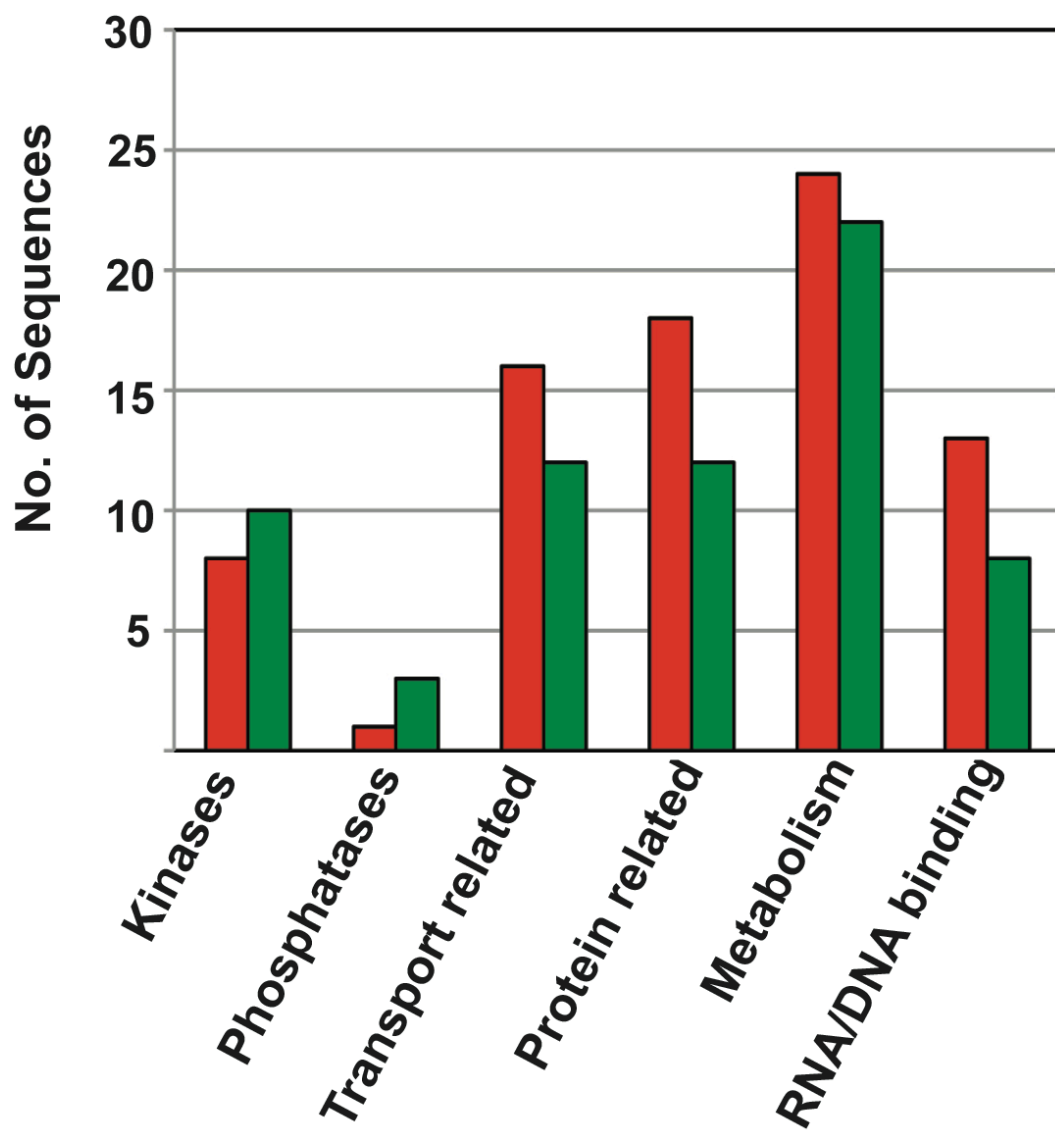


Figure 4.5. Daily variation of kinase activity and their efficiency in *Lingulodinium*

A Group-based Prediction System was used to identify the most likely kinases responsible for phosphorylation of all the 673 different serine/threonine phosphosites. In (A), the kinase categories responsible for phosphorylation at sites in hyperphosphorylated peptides at either ZT2 (red) or ZT14 (green) are shown along with the sites with unchanged levels of phosphorylation (blue). In (B), the efficiency of the different kinase classes at the two times was estimated by comparing the number of kinases in each group with the number of phosphosites assigned to each group. A straight line through the origin would be expected for a direct correlation between abundance and activity. The ‘Others’ category consists of kinases other than that of CK2. Details of the kinase classes are provided in Figure 4.S1.

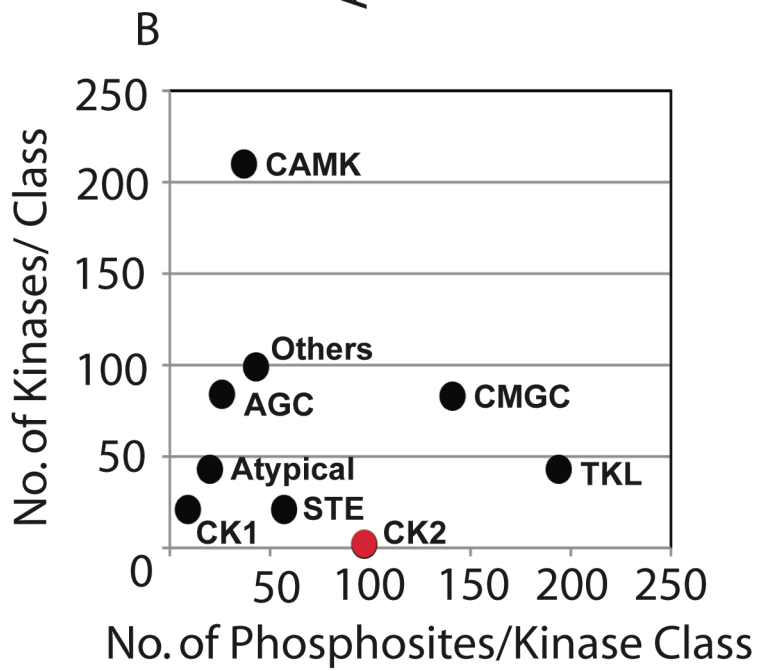
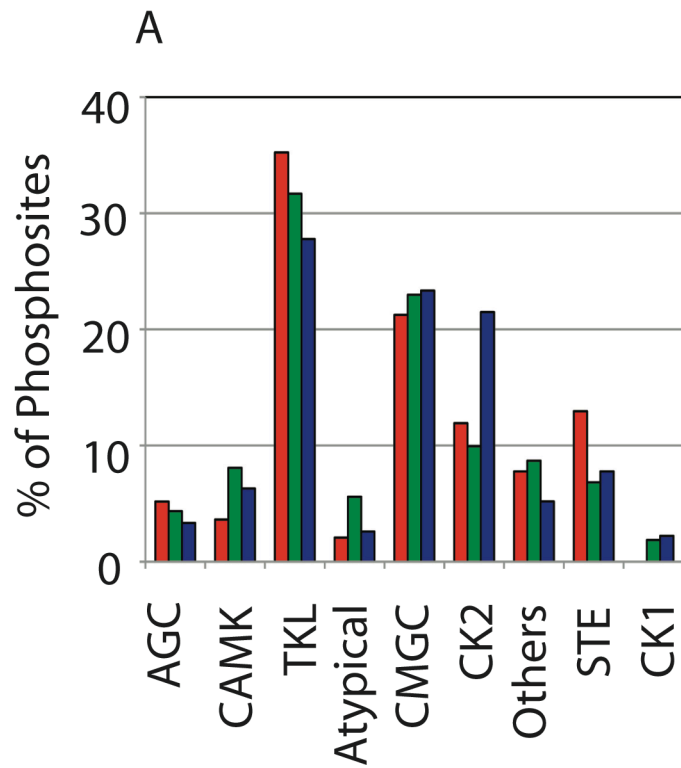
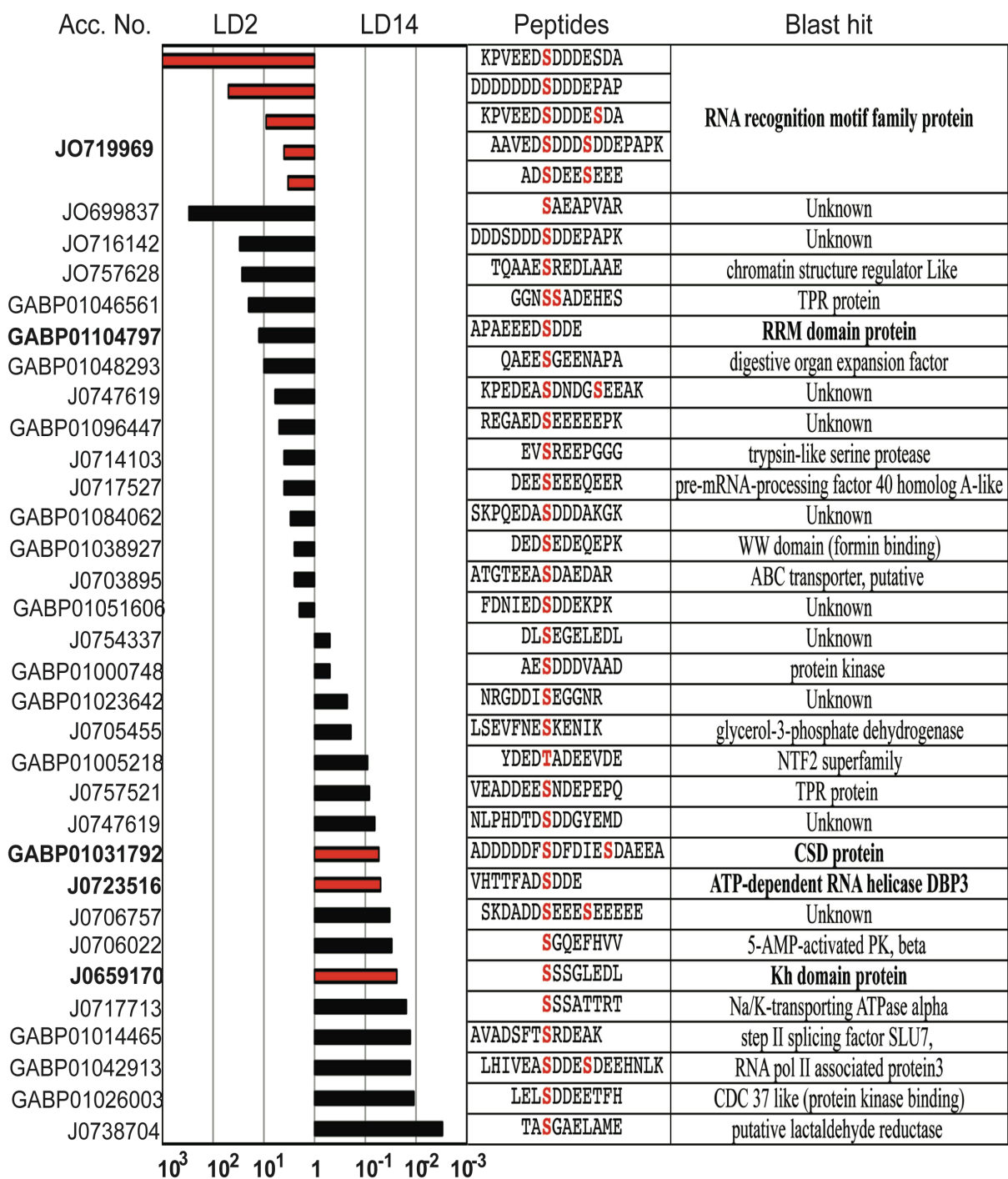


Figure 4.6. Many RNA binding proteins are among the predicted CK2 targets

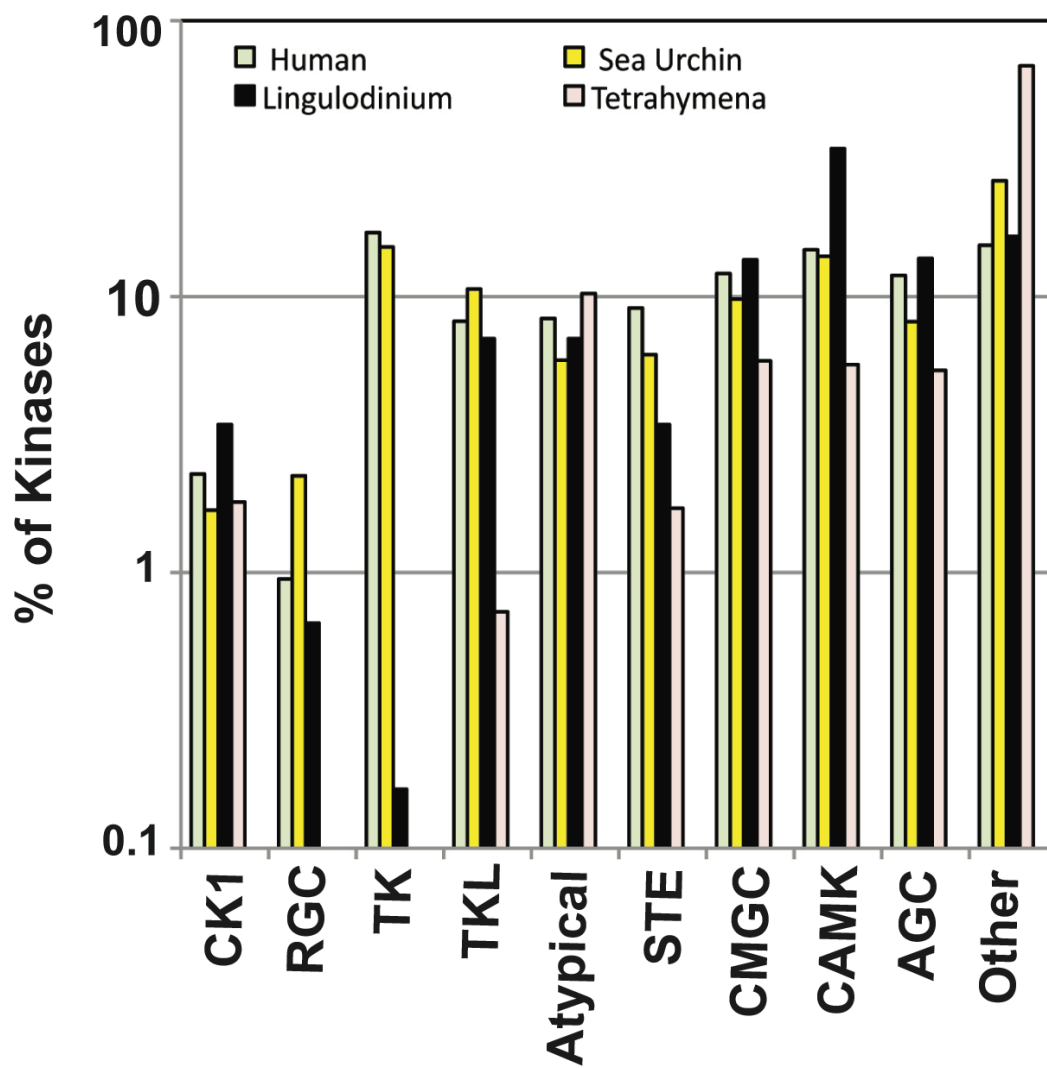
Potential CK2 substrates hyperphosphorylated (≥ 2 -fold) at ZT2 or ZT14 are shown together with the observed intensity change (shown on a Log scale). The RNA binding proteins are in bold and their fold change is shown as red bars. The phosphorylated serine/threonine residues within the peptide sequences are shown in red. The proteins are ordered to show maximum and minimum phosphorylation intensity at ZT2 at the top and bottom of the table, respectively. Note that the first five peptides are derived from the same protein.



Supplementary Figure Legends

Figure 4.S1. Comparison of kinases

Description: A comparison of the kinome of *Lingulodinium* with that of three other eukaryotes, including *Tetrahymena*, a ciliate phylogenetically related to the dinoflagellates. The number of kinases (% of total) is shown on the y-axis using a logarithmic scale. The abbreviations used along the X-axis correspond to the well-known kinase classes. A total of 528, 357, 1111 and 611 kinases were found in Human, Sea urchin, *Tetrahymena* and *Lingulodinium*, respectively. The kinase classes are CK1, Casein Kinase1; RGC, Receptor Guanylate Cyclases; TK, Tyrosine Kinase; TKL, Tyrosine Kinase Like; CAMK, Calcium and Calmodulin-regulated kinases; CMGC, contains CDK, MAPK, GSK3 and CLK kinases; STE, contains the homologs of yeast Sterile 7, Sterile 11, and Sterile 20 kinases; AGC, contains PKA, PKG and PKC.



Supplementary Tables

Table 4.ST1. *Lingulodinium* kinases

Description: The different types of *Lingulodinium* kinases. The kinases with proven role in eukaryotic circadian biology are marked in blue. The kinase sub-group with maximum representation is in red (almost three quarter of the 200 CAMKs are within this subgroup). The details of the kinase classes are described in Figure 4.S1.

| AGC | CK-1 | Atypical | CMGC | STE | TKL | Other | CAMK | TK |
|----------|--------------|-------------|-------------|-------------------------------|----------------|------------------------------|------------|---------|
| PKG | CK1-A | PIKK/FRAP | CDK/CDC2 | STE11/STE-11- Unclassified | LRRK | NAK | CAMK1 | SRC/FRK |
| PKA | CK1-G | PIKK/ATM | CDK/CDK5 | STE11/MEKK1 | IRAK | NAK/GAK | CAMK2 | |
| PKC/IOTA | CK1- D/E | PIKK/ATR | CDK/CDK7 | STE11/MEKK2 | TKL-UNIQUE | NEK/NEK1 | CAMK-Tt | |
| RSK | CK1- TTBK | ABC1 | CDKL | STE7/STE-7- Unclassified | TKL- CILIAE | NEK/NEK4 | CDPK | |
| NDR | | ABC1-B | CLK | STE/DICTY 2 | MLK/MLK | NEK/NEK6 | CAMKL/MARK | |
| SGK | | ABC1/ABC1-C | MAPK/ERK | STE20/FRAY | MLK/HH498 | NEK/NEK9 | CMAKL/QIK | |
| Akt | | RIO/RIO-2 | MAPK/ERK7 | STE20/YSK | DICTY4/SWR | NEK/NEK- Unclassified | CAMKL/NuaK | |
| | | RIO/RIO-1 | DYRK/PRP4 | STE20/MST | DICTY4/DRK | CK2 | CAMKL/AMPK | |
| | | PDHK/PDHK | DYRK/DYRK P | STE20/NINAC | | WNK | RAD53 | |
| | | ALPHA/VWL | DYRK/YAK | | | ULK/ULK | DCAMKL | |
| | | | DYRK/DYRK 2 | | | ULK/FUSED | DAPK/DAPK | |
| | | | RCK/MOK | | | IKS | MAPKAP/MNK | |
| | | | RCK/MAK | | | TTK | PHK | |
| | | | GSK | | | WEE | | |
| | | | SRPK | | | AUR | | |
| | | | | | | TLK | | |
| | | | | | | SCY1 | | |
| | | | | | | Ciliate-E2/Ciliate- E2B | | |
| | | | | | | CAMKK/CAMKK- Unclassified | | |

Table 4.ST2. The identification of proteins containing the 527 phosphopeptides in ZT2 and ZT14 extracts of *Lingulodinium*

| LD2 | LD14 | UniProt ID | Peptide Sequence | Pep Modification | best blast hit description | Peptide Score |
|-------|-------|--------------|---------------------------|---|---|---------------|
| 9E+07 | 1E+04 | GABP01057639 | RPRRPSSLPK | Phospho (ST)-S6 | Unknown | 32.86 |
| 4E+07 | 1E+04 | GABP01093776 | KTFEATSAASSRPGK | Phospho (ST)-T2 Phospho (ST)-S7 | Proteasome 26S non-atpase subunit | 27.53 |
| 3E+07 | 1E+04 | GABP01065238 | SPLPGAWTSRSTSPVSGGTSAASPR | Phospho (ST)-S17 | Unknown | 34.43 |
| 3E+07 | 5E+05 | GABP01084233 | TSPCCTVARSTACTPAEAVR | Phospho (ST)-S2 Carboxymethyl (C)-C4 | Unknown | 29.55 |
| 2E+07 | 1E+07 | JO754267 | RPGPHAVGRPRQTQGR | Phospho (ST)-T12 Deamidated [474]-Q13 | Protein kinase domain containing protein | 21.31 |
| 2E+07 | 1E+04 | GABP01003046 | ASLLRQIK | Phospho (ST)-S2 | Unknown | 20.02 |
| 2E+07 | 1E+07 | JO719969 | RPAPEDSDDDEEPPAK(K) | Phospho (ST)-S7 | RNA recognition motif family protein | 109.91 |
| 1E+07 | 1E+04 | GABP01037922 | AAAPTALGR | Phospho (ST)-T5 | Zf-dhbc type | 20.19 |
| 1E+07 | 3E+05 | GABP01075942 | KAITATITIWLL | Phospho (ST)-T8 Deamidated [474]-N11 | VAPB | 23.26 |
| 1E+07 | 1E+04 | JO719969 | KPVEEDSDDDESDA | Phospho (ST)-S7 | RNA recognition motif family protein | 73.55 |
| 1E+07 | 1E+04 | GABP01024383 | LGSVAQGSFK | Deamidated [474]-Q6 Phospho (ST)-S8 | Pyruvate ferredoxin oxidoreductase | 28.47 |
| 1E+07 | 1E+04 | JO722903 | VSSRASSPR | Phospho (ST)-S3 | methionine S-adenosyl transferase | 15.29 |
| 9E+06 | 5E+06 | GABP01049965 | SLATSNPGESPQGPQPK | Phospho (ST)-S10 | Pumilio domain containing pentatricopeptide repeat-containing protein | 67.07 |
| 9E+06 | 4E+05 | JO692895 | SLQGVKLNPK | Phospho (ST)-S1 Deamidated [474]-N8 | | 43.3 |
| 8E+06 | 2E+06 | JO700473 | NATALLGNSGVQ | Deamidated [474]-N1 Deamidated [474]-N8 Phospho (ST)-S9 | salivary gland secretion 1 | 28.55 |
| 8E+06 | 1E+04 | GABP01060748 | VAVHGPALGTARHR | Phospho (ST)-T11 | Unknown | 16.91 |
| 7E+06 | 8E+06 | GABP01033861 | GGRTCHAAIAR | Phospho (ST)-T4 | Phosphoenolpyruvate synthase | 39.26 |
| 7E+06 | 4E+05 | GABP01018185 | CDFTLLDGGDLLGTTTEMLNLNAP | Phospho (ST)-T4 Phospho (ST)-S29 | hypothetical protein | 21.94 |
| 7E+06 | 1E+04 | GABP01032640 | GHPCCDQVGAPTLPGIDTPSIALR | Deamidated [474]-Q7 Phospho (ST)-T12 | Unknown | 18.25 |
| 6E+06 | 1E+04 | GABP01008467 | EVSGEAQTAPTPLSAQTAR | | eukaryotic translation initiation factor 3 delta subunit | 16.07 |
| 6E+06 | 1E+04 | GABP01044390 | LGISPASPTPTTQ | Phospho (ST)-S4 | Glycogen phosphorylase family | 20.76 |
| 6E+06 | 2E+06 | GABP01114065 | QGLRPAPSGIGGSTRPCSSRSCGR | Phospho (ST)-S8-S14-S22 | farnesyltransferase/geranylgeranyltransferase type 1 alpha subunit | 15.01 |
| 6E+06 | 4E+06 | GABP01048268 | TQSCRWPWPCGAPSTCARWTCSR | Phospho (ST)-T15 Carboxymethyl (C)-C21 | Unknown | 18.62 |
| 5E+06 | 1E+04 | GABP01023398 | ALGGPQRGSRPCPRPR | Deamidated [474]-Q6 Phospho (ST)-S9 | Homocysteine S Methyltransferase protein | 26.13 |
| 5E+06 | 2E+06 | JO703895 | ATGTEEASDAEDAR | Phospho (ST)-S8 | ABC transporter | 60.58 |
| 5E+06 | 4E+05 | GABP01104797 | KPAPAEEDSDDE | Phospho (ST)-S10 | RRM domain protein | 86.07 |
| 5E+06 | 3E+06 | GABP01112644 | SFATLQPSTYGDSDL | Deamidated [474]-Q6 Phospho (ST)-S13 | Unknown | 17.05 |
| 5E+06 | 6E+06 | GABP01025983 | SLCCIQSSNDPWNLPLPK | Phospho (ST)-S1-S8 | Unknown | 25.83 |
| 5E+06 | 1E+07 | JO759265 | GVLPLVTQSFVGTDSVIAK | Deamidated [474]-Q8 Phospho (ST)-S9 | Pyruvate kinase, putative | 108.85 |

| | | | | | | |
|-------|-------|--------------|---|--|--|--------|
| 5E+06 | 3E+06 | JO733631 | VGDEQGPSVLDA GPNYDSEETAK K | Phospho (ST)-S19 | conserved hypothetical protein | 79.7 |
| 5E+06 | 4E+06 | GABP01042453 | FLQLLAVLSLCQPR | Phospho (ST)-S9 | Unknown | 24.46 |
| 4E+06 | 3E+06 | GABP01104797 | KQTFSDSDEEEKPPAK | Phospho (ST)-S5 Phospho (ST)-S7 | RRM domain containing protein RNA recognition motif. | 51.97 |
| 4E+06 | 3E+06 | JO719969 | QTFNDSDDDEEDDEPAPK | Phospho (ST)-S6 Deamidated [474]- | family protein | 82.86 |
| 4E+06 | 8E+05 | JO736363 | AASSSCNRSTS QDCAGLV SGLGPGPE GATVTSR | Q12 Phospho (ST)-S19- T29-T31 | Unknown | 31.92 |
| 4E+06 | 1E+04 | GABP01111881 | LLATLLVVDLGR | Phospho (ST)-T4 | hypothetical protein | 17.36 |
| 4E+06 | 2E+05 | GABP01097215 | PSWELLKLLTSDTTELVSALATG R | Phospho (ST)-S2-T10-S20- T25 | Unknown | 19.99 |
| 4E+06 | 1E+04 | JO694135 | SQDGS HCSGAVPGMRATR | Phospho (ST)- S1 Deamidated [474]- Q2 Phospho (ST)-T17 | Unknown | 33.48 |
| 4E+06 | 1E+06 | JO704770 | TAALASTVPESARMNSWYCSTLPHS VSR | Oxidation (M)- M14 Phospho (ST)-S20- T21-S25 | CG7139, isoform A | 17.26 |
| 4E+06 | 1E+04 | JO748208 | VTSSAGR NHTGLFLMSSH | Deamidated [474]- N8 Phospho (ST)- T10 Phospho (ST)-S18 | heat shock protein 90 DNAJ N-terminal domain- containing protein | 17.79 |
| 3E+06 | 3E+06 | GABP01015275 | RGDDSGGEEEQFR | Phospho (ST)-S5 | Unknown | 47.65 |
| 3E+06 | 1E+06 | GABP01084062 | AEEEDPPSKPQEDASDDDAKGK | Phospho (ST)-S15 | NOT2/NOT3/NOT5 domain- containing protein | 45.16 |
| 3E+06 | 1E+06 | JO720544 | AVGTPQNAASGGGSATK | Phospho (ST)-T4 | Bifunctional purine biosynthesis | 125.79 |
| 3E+06 | 1E+04 | GABP01080240 | ELTQATGQPAAASF | Phospho (ST)-T6 Deamidated [474]- Q4 Oxidation (M)- | LRR protein | 32.54 |
| 3E+06 | 1E+04 | GABP01026460 | ESVQAGMRSEQLGLK | M7 Phospho (ST)-S9 Oxidation (M)-M5 Phospho (ST)-T6 Phospho (ST)-T14 | Lysine decarboxylase domain containing protein | 19.88 |
| 3E+06 | 8E+05 | JO748411 | GLLAGQGGGLLVLLSGRDR | Phospho (ST)-S15 | Unknown | 18.49 |
| 3E+06 | 1E+04 | GABP01055215 | HAASVWKMASNAACSK | Phospho (ST)-S4 | type I polyketide synthase- like protein | 15.04 |
| 3E+06 | 1E+04 | GABP01042469 | KPASGPGRATFR | Phospho (ST)-T11 | n-adenine specific methyltransferase 1 | 20.81 |
| 3E+06 | 1E+04 | JO740256 | LESTLHLVLVGR | Phospho (ST)-T4 | Chloroplast o methyltransferase | 16.14 |
| 3E+06 | 1E+04 | GABP01046551 | QLVSMEAVLR | Phospho (ST)-S4 Phospho (ST)-T12 Phospho (ST)-S13 Deamidated | Unknown | 41.05 |
| 3E+06 | 1E+04 | JO707803 | RQAPQTSGSAGTSAPGPCNR | [474]-N19 | Unknown | 17.97 |
| 3E+06 | 1E+04 | JO699837 | SAEAPVARAQGPGR | Phospho (ST)-S1 Oxidation (M)-M2 Phospho (ST)-S14 | Unknown | 15.23 |
| 3E+06 | 1E+04 | GABP01035379 | SMLSECFRPTAPSSR | Phospho (ST)-T13 | MAPK | 19.4 |
| 3E+06 | 1E+04 | GABP01091546 | SRASWITSITPTR | Phospho (ST)-S3 | carbohydrate binding protein | 16.62 |
| 3E+06 | 4E+05 | GABP01012985 | STSNPYDERPVGK | Phospho (ST)-T7 | calcium-binding protein | 26.16 |
| 3E+06 | 1E+04 | GABP01028663 | VEPLLTAGK | Phospho (Y)-Y1-S5-S9 | peptidase C14 | 21.82 |
| 3E+06 | 1E+06 | GABP01010423 | YTPSNRQSTVQPASQR | Carboxymethyl (C)- C7 Phospho (ST)-S16 | Unknown | 22.62 |
| 3E+06 | 2E+06 | GABP01024322 | TTARLVCAGPLALLPSASSPVAR | DDE superfamily endonuclease | cAMP-dependent protein | 15.37 |
| 3E+06 | 5E+06 | GABP01036663 | TYTLCGTPEYIAPEVLLNK | Phospho (ST)-T1 | kinase catalytic subunit | 97.02 |
| 3E+06 | 3E+06 | JO744240 | EPLLLVGSFNSWSVEKAR | Phospho (ST)-S13 | Unknown | 19.14 |
| 2E+06 | 4E+06 | GABP01036474 | PSVIAAGSSRR | Phospho (ST)-S9 | Unknown | 27.42 |

| | | | | | | |
|-------|-------|--------------|-----------------------------|---|--|--------|
| 2E+06 | 3E+06 | GABP01024922 | ALSPAAAAGEGPQGR | Phospho (ST)-S3 | pinin/sdk/mema domain containing protein | 73.02 |
| 2E+06 | 1E+06 | GABP01028776 | AVTEPVPTQQASPDASPTK | Phospho (ST)-S16 | Arf-GAP like | 64.39 |
| 2E+06 | 3E+06 | GABP01051799 | LDSNEDVRPSAGMR | Phospho (ST)-S3 | Unknown | 35.01 |
| 2E+06 | 2E+06 | JO720445 | LLSRDEESGPTAK | Phospho (ST)-S3 | predicted protein | 23.59 |
| 2E+06 | 1E+06 | GABP01051835 | APSPAPAAPAQAR | Phospho (ST)-S3 | Forkhead associated domain containing | 56.07 |
| 2E+06 | 7E+05 | GABP01055210 | KDSANLVCAVMGQMDPFGWGGLCQ | Phospho (ST)-S3 Oxidation (M)-M14 Phospho (ST)-S29 | type I polyketide synthase-like protein | 19.67 |
| 2E+06 | 7E+05 | GABP01024454 | KPGEDGSPHSLQEYSK | Phospho (ST)-S8 | Unknown | 70.79 |
| 2E+06 | 3E+05 | JO735126 | LEPSSLPQLK | Phospho (ST)-S4 | type I polyketide synthase-like protein | 20.59 |
| 2E+06 | 1E+04 | GABP01003193 | LLPSVRSAGQTWQR | Phospho (ST)-S4 Deamidated [474]-Q13 | TRP protein | 18.86 |
| 2E+06 | 1E+04 | GABP01030993 | LRTLGTTPCSQTPASR | Carboxymethyl (C)-C9 Phospho (ST)-S15 | ZN finger RANBP2 type | 25.44 |
| 2E+06 | 1E+04 | GABP01004864 | QDVKAISQSVQQLR | Phospho (ST)-S9 Deamidated [474]-Q11 | GRIP and coiled-coil domain containing | 33.32 |
| 2E+06 | 1E+04 | GABP01074973 | RGQGQVVGASSGGLRPRPHLAR | Phospho (ST)-S10 | Unknown | 21.82 |
| 2E+06 | 1E+06 | JO742036 | RNSISATPVSNDR | Phospho (ST)-S3 | cAMP-dependent protein kinase regulatory subunit | 56.03 |
| 2E+06 | 6E+05 | JO701179 | RSDVPGLASAGASTLMSPCGPCTEK | Phospho (ST)-S9 Oxidation (M)-M16 Carboxymethyl (C)-C22 | Unknown | 15.5 |
| 2E+06 | 1E+06 | GABP01082788 | SGTLGLQGPDSGASAGR | Phospho (ST)-S11 | Unknown | 41.1 |
| 2E+06 | 2E+06 | GABP01091939 | SLYAQLSSSAVCSLVSRTGQVSHAMP | Phospho (ST)-S16 Deamidated [474]-Q20 | Unknown | 18.41 |
| 2E+06 | 9E+04 | GABP01064394 | SPAEGAASIPSRAPAALAK | Phospho (ST)-S1 | PK domain containing protein | 27.23 |
| 2E+06 | 1E+04 | GABP01020980 | SRPVALPPRPACPGMAK | Phospho (ST)-S1 Carboxymethyl (C)-C13 | argininosuccinate lyase | 17.44 |
| 2E+06 | 1E+04 | GABP01078536 | SSRASGSPPAPAQTSPPR | Phospho (ST)-S7-T14-S15 | SPRY domain containing protein | 17.72 |
| 2E+06 | 1E+04 | GABP01027148 | TASLVSVTPSSACVSRSLPCGMGVTPS | Phospho (ST)-T1-S3-S6-S26-S27 | Glucosidase | 20.44 |
| 2E+06 | 1E+05 | GABP01098060 | TNGHVLVPSYQLQHWR | Phospho (ST)-T1 Deamidated [474]-N2 Phospho (ST)-S9 | Unknown | 15.74 |
| 2E+06 | 1E+04 | GABP01078147 | TSLALTLPIAPR | Phospho (ST)-S2 | Unknown | 20.92 |
| 2E+06 | 2E+06 | JO765479 | RPLLVTAAATLLNFCSLCLMPPTTK | Phospho (ST)-S15 Phospho (ST)-T22 | Unknown | 28.44 |
| 2E+06 | 1E+06 | GABP01031954 | ALRPEAGSGPGSPSR | Phospho (ST)-S12 | Unknown | 23.05 |
| 2E+06 | 1E+06 | GABP01086664 | LENVMVDMESPKR | Phospho (ST)-S10 | Protein kinase domain containing protein | 36.88 |
| 2E+06 | 1E+06 | GABP01084062 | ARAEEDPPSKPQEDASDDDAK | Phospho (ST)-S17 | Unknown | 42.7 |
| 2E+06 | 1E+06 | GABP01078701 | YSVGKPEDQEGPAVLDSGDPNYDSE | Phospho (ST)-S24 | Unknown | 132.79 |
| 2E+06 | 3E+06 | GABP01051532 | GEEEAQETTETK | Phospho (ST)-T11 | Unknown | 20.19 |
| 2E+06 | 2E+06 | GABP01108400 | CLYTRLAVQATQK | Phospho (ST)-T21 Deamidated [474]-Q27 | Unknown | 19.96 |
| 2E+06 | 1E+06 | GABP01035382 | ALGMAPGEVSPIEHLFGPDTPAEEL | Phospho (ST)-S8 | MAPK2 | 59.39 |
| 2E+06 | 1E+06 | JO758863 | QVVSLLWAP | Phospho (ST)-S3 | Nucleolar protein Nop56, calcium-dependent protein | 116.44 |
| 2E+06 | 1E+06 | JO748908 | EDGSDESPVPVAPHTQR | Phospho (ST)-S12 | kinase | 64.05 |
| 2E+06 | 2E+06 | GABP01012072 | HLQSADAEPDGSPVR | Deamidated [474]-N2 Phospho (ST)-S9 Oxidation (M)-M12 | Unknown | 24.86 |
| 1E+06 | 2E+06 | GABP01025477 | LNALLPAVSRPMAPSGSEVIRTFAAH | Carboxymethyl (C)- | Unknown | 16.5 |
| | | | AK | | | |
| | | | PACTAAQAASAQPPRRPTLLSFAGA | | | |

| MAAK | | | | C3 Phospho (ST)-T4 | | |
|-------|-------|--------------|---------------------------|---|---|--------|
| 1E+06 | 9E+05 | GABP01049965 | TVSGADASSEPSAEGGSPMK | Phospho (ST)-S17 | Pumilio domain containing | 57.81 |
| 1E+06 | 3E+06 | GABP01015624 | ASLLEEAEDESEDEPEK(GK) | Phospho (ST)-S10 | Splicing factor 3B subunit | 108.81 |
| 1E+06 | 7E+05 | GABP01068216 | FDHIEDSDDETPAAKPVPK | Phospho (ST)-S7 | Unknown | 35.72 |
| 1E+06 | 5E+06 | JO704636 | VVLAYEPVWAIGTGK | Phospho (ST)-T13 | triosephosphate isomerase | 69.48 |
| 1E+06 | 9E+05 | GABP01071379 | GVLPLVTQSFVGTDSVIAK | Phospho (ST)-S9 | Pyruvate kinase | 92.32 |
| 1E+06 | 1E+06 | JO693972 | RTASNPAGMLLLMPSLK | Phospho (ST)-S4 Oxidation (M)-M13 | Memo (mediator of ErbB2-driven cell motility) | 19.51 |
| 1E+06 | 1E+06 | GABP01104797 | AKKPAPAEEDSDDE | Phospho (ST)-S12 | RRM domain containing protein | 59.45 |
| 1E+06 | 2E+06 | JO752062 | ICGSKPISNIR | Phospho (ST)-S4 Phospho (ST)-S8 Deamidated [474]-N9 | DEAD-box ATP-dependent RNA helicase | 23.4 |
| 1E+06 | 2E+06 | JO708271 | AGSITDDVFNMVDR | Phospho (ST)-S3 | ALVEOLIN1 | 85.49 |
| 1E+06 | 8E+05 | JO702217 | EEEKEESEEPPAPNAELTK | Phospho (ST)-S7 | hypothetical protein | 63.91 |
| 1E+06 | 7E+05 | GABP01007088 | QLYELEMEQK | Deamidated [474]-Q1 Phospho (Y)-Y3 Carboxymethyl (C)- | Leo1p | 17.5 |
| 1E+06 | 8E+04 | GABP01054884 | AMRLSCSSHQALR | C6 Phospho (ST)-S8 | dynein heavy chain | 15.69 |
| 1E+06 | 1E+06 | GABP01005490 | PELYEQLSK | Phospho (ST)-S21 | SART-1 FAMILY PROTEIN | 62.87 |
| 1E+06 | 6E+05 | GABP01051606 | FDNIEDSDDEKPK | Phospho (ST)-S7 | Unknown | 38.41 |
| 1E+06 | 5E+05 | GABP01024074 | GAEVDDSPPR | Phospho (ST)-S7 | Unknown | 29.94 |
| 1E+06 | 1E+04 | GABP01088221 | GGVSRSLGTVPLSPVSAK | Phospho (ST)-S6 | nose resistant to fluoxetine protein 6-like | 20.01 |
| 1E+06 | 4E+05 | JO709548 | GHAGPSAVAQQHTPPSHACLHLLAA | Phospho (ST)-S6-T13-S16 | RRM domain protein | 16.92 |
| 1E+06 | 1E+04 | GABP01014219 | GSVTDLIALSALK | Phospho (ST)-S10 | Hid-1 like | 18.22 |
| 1E+06 | 5E+05 | JO712444 | HGLEGPPWHRALLLTPPAR | Phospho (ST)-T15 Phospho (ST)-T16 | hypothetical protein | 23.27 |
| 1E+06 | 1E+04 | GABP01040404 | LEVHSALK | Phospho (ST)-S5 | Unknown | 29.7 |
| 1E+06 | 1E+04 | GABP01066742 | LVAVPGSRAASR | Phospho (ST)-S7 | HSP | 17.86 |
| 1E+06 | 1E+04 | GABP01025294 | NLALVFPSLSSR | Deamidated [474]-N1 Phospho (ST)-S11 | RNA polymerase II associated protein | 26.28 |
| 1E+06 | 2E+05 | JO757528 | QTVGLLQRSHGQITSK | Phospho (ST)-S15 | SIT4 phosphatase-associated family protein | 18.27 |
| 1E+06 | 6E+05 | GABP01060639 | RAIHGVPVAPQSIPLEP | Phospho (ST)-S12 | Unknown | 15.91 |
| 1E+06 | 4E+05 | GABP01096447 | REGAEDSEEEEPK | Phospho (ST)-S7 | Unknown | 86.13 |
| 1E+06 | 1E+04 | GABP01012171 | RPSGLNSNSKNSCPNLPLWPT | Phospho (ST)-S7-S9-S12 | mago nashi protein | 19.67 |
| 1E+06 | 4E+05 | JO695593 | STAMSPEKIEGR | Phospho (ST)-S5 | hypothetical protein | 24.26 |
| 1E+06 | 1E+04 | GABP01026990 | YRMVICSVLAGK | Phospho (Y)-Y1 Oxidation (M)-M3 Carboxymethyl (C)-C6 | Unknown | 18.01 |
| 1E+06 | 1E+06 | JO706453 | WGAVNPGSPGGAGGWR | Phospho (ST)-S8 | fibrinogen A-alpha chain | 47.91 |
| 1E+06 | 1E+06 | GABP01086664 | LIDFDTMQDWEPPSPK | Phospho (ST)-S14 | Protein kinase domain containing protein | 69.59 |
| 9E+05 | 1E+06 | GABP01086664 | ADIFDDLPGENWVGSPAMR | Phospho (ST)-S15 | Protein kinase domain containing protein | 37.32 |
| 9E+05 | 6E+05 | JO743538 | LEPEQSGSSPPAR | Phospho (ST)-S9 | serine/threonine protein phosphatase | 20.79 |
| 9E+05 | 4E+05 | JO695062 | GVLPLVTQSFVGTDSVIAK | Phospho (ST)-S9 | Pyruvate kinase | 97.65 |
| 9E+05 | 1E+05 | JO719969 | KPVEEDSDDDESDA | Phospho (ST)-S7 Phospho (ST)-S12 | RNA recognition motif. family protein | 40.54 |
| 9E+05 | 3E+05 | GABP01059430 | LLGGSLAGLYRPGAGSK | Phospho (ST)-S16 | 2,5 didehydrogluconate reductase | 15.5 |
| 9E+05 | 1E+04 | GABP01099849 | NPDTKAAAAVTSLVGVTGAVSSV | Deamidated [474]-N1 Phospho (ST)-S24 | Unknown | 15.95 |
| 9E+05 | 1E+04 | GABP01099849 | DK | | | |

| | | | | | | |
|-------|-------|--------------|---------------------------|---|--|-------|
| 9E+05 | 4E+05 | GABP01038668 | SGPQTVVCPCGSKTMTQAPR | Carboxymethyl (C)- C10(Oxidation (M)- M15 Phospho (ST)-T16 | hect E3 ubiquitin ligase S-adenosyl-homocysteine hydrolase | 15.28 |
| 9E+05 | 1E+04 | GABP01099732 | SLPRAVNCLTPLTMLSSR | Phospho (ST)-S18 | | 16.04 |
| 9E+05 | 6E+05 | GABP01092558 | GPASDLELADSSGR | Phospho (ST)-S11 | DnaJ domain containing Zn-finger RING-CH type domain | 16.78 |
| 9E+05 | 7E+05 | GABP01034357 | RQISSPSTSIAPAPVR | Phospho (ST)-S10 | | 24.29 |
| 9E+05 | 8E+05 | GABP01049965 | DRAPTPEDASEEANVNR | Phospho (ST)-T5 | Pumilio domain containing | 35.49 |
| 9E+05 | 1E+06 | GABP01013062 | AGIEDGDEQDAEAAAHGGDQRPVSR | Phospho (ST)-S24 | Unknown Chalcone and Stilbene domain containing | 85.29 |
| 8E+05 | 1E+06 | GABP01078304 | LGSLNPSITPALLEASEALPK | Phospho (ST)-S16 Deamidated [474]- Q1 Phospho (ST)- S7 Deamidated [474]-N9 | Cold Shock domain containing protein | 33.91 |
| 8E+05 | 5E+05 | GABP01025605 | QCGTITSWNMARAGK | Phospho (ST)-S4 | | 29.36 |
| 8E+05 | 1E+06 | GABP01001006 | RPDSPGLQDR | Phospho (ST)-T1 Phospho (ST)-S8 | Unknown | 37.44 |
| 8E+05 | 6E+05 | GABP01059771 | TPASLAGSPVPVPLGASR | Phospho (ST)- S4 Deamidated [474]-Q5 Phospho (ST)-T6 Phospho (ST)-S17 | Unknown hypothetical protein atp dependent RNA helicase ddx family RNA polymerase II associated protein 3 | 26.62 |
| 8E+05 | 4E+05 | JO759786 | ARGSQLPL | | | 15.56 |
| 8E+05 | 2E+05 | GABP01091378 | EGDVITCLLDRENQTISYCK | | | 20.7 |
| 8E+05 | 3E+05 | GABP01042912 | EPVPAPETTTANESPK | Phospho (ST)-S15 | | 40.27 |
| 8E+05 | 2E+05 | JO714103 | EVSREEPGGGPQSR | Phospho (ST)-S3 Deamidated [474]- Q9 Phospho (ST)-S16 | hypothetical protein 3 isopropylmalate dehydratase | 22.42 |
| 8E+05 | 1E+04 | GABP01043956 | GDLLLPQQAPGPEGSR | Phospho (ST)-T8 | | 19.09 |
| 8E+05 | 1E+04 | JO716193 | GGAVRGGTK | Phospho (ST)-S4 | hypothetical protein RAP protein | 22.29 |
| 8E+05 | 1E+04 | JO763941 | LSASTRSGSR | Phospho (ST)-S6 | | 16.84 |
| 8E+05 | 1E+04 | GABP01081971 | NFLGQSGAADLLTYAASLGCAR | | Unknown t complex protein 1 subunit alpha | 31.2 |
| 8E+05 | 1E+04 | GABP01084658 | WSGAIDGIR | Phospho (ST)-S2 | | 33.36 |
| 8E+05 | 1E+06 | GABP01094643 | WDEVESDPDEPCPDVAK | Phospho (ST)-S6 | Unknown | 92.5 |
| 8E+05 | 9E+05 | GABP01016050 | LPIQRTLYIVGLSR | Phospho (ST)-S13 | type I fatty acid synthase | 24.08 |
| 7E+05 | 1E+06 | JO756760 | TYTLCGTPEYIAPEVLLNK | Phospho (ST)-T3 | protein kinase, putative | 44.37 |
| 7E+05 | 6E+05 | JO755585 | LLPEAGESPTTK | Phospho (ST)-S8 | Unknown RNA binding protein NOVA- 2 | 16.8 |
| 7E+05 | 4E+05 | GABP01029422 | RSSPRPSLAPVRPRPAGADGRP | Phospho (ST)-S2 | | 16.6 |
| 7E+05 | 7E+05 | JO715474 | MRPVPDPLLPESELDSPVSSAK | Phospho (ST)-S17 Phospho (ST)-S7 Phospho (ST)-T11 | hypothetical protein Unknown | 58.69 |
| 7E+05 | 4E+05 | GABP01068216 | FDHIEDSDDETPAAKPVPK | Phospho (ST)-T4 Oxidation (M)-M13 Deamidated [474]-Q16 | nucleoside diphosphate sugar epimerase | 82.44 |
| 7E+05 | 3E+05 | GABP01020442 | VLVTGAGGRTGSMVLQK | Phospho (ST)-S13 | | 23.61 |
| 7E+05 | 6E+05 | GABP01047669 | AVAAVDAGQPPDSPAQR | Deamidated [474]- Q5 Phospho (ST)-T8 | Unknown Serine/threonine kinase domain protein tetrahydrofolate dehydrogenase | 22.95 |
| 7E+05 | 1E+06 | GABP01034207 | SSVQQAFTTR | | | 23.32 |
| 7E+05 | 1E+07 | GABP01073223 | ARPPPAASPASHLR | Phospho (ST)-S11 Phospho (ST)-S8 Phospho (ST)-S10 | | 15.2 |
| 6E+05 | 6E+05 | GABP01045657 | RPSGQAPSPSPAR | | Unknown calcium-dependent protein kinase | 44.29 |
| 6E+05 | 4E+05 | JO732494 | LTSSFTLTLEAPSK | Phospho (ST)-S12 Carboxymethyl (C)- C6 Phospho (ST)- S7 Phospho (ST)-S8 | | 15.67 |
| 6E+05 | 1E+04 | GABP01048213 | CSERVCSVVVPAMR | | Resistance nodulation cell division superfamily | 18.61 |
| 6E+05 | 1E+04 | GABP01045041 | LCVSASTRSSR | Phospho (ST)-S10 | Myosin XI-I | 19.04 |

| | | | | | | |
|-------|-------|--------------|-----------------------------|--|--|--------|
| 6E+05 | 2E+05 | JO754791 | LSFMLK | Phospho (ST)-S2 | cytosolic tRNA-Ala synthetase | 16.51 |
| 6E+05 | 1E+05 | JO742036 | RNSISATPVSNDR | Phospho (ST)-S3 Phospho (ST)-T7 | cAMP-dependent protein kinase regulatory subunit | 41 |
| 6E+05 | 3E+05 | GABP01082788 | SGTLGLQGPDSPGASAGR | Phospho (ST)-T3 | Unknown | 49.77 |
| 6E+05 | 3E+05 | JO718357 | LRSCMHVLLK | Phospho (ST)-S3 | ADP-ribosylation factor, putative | 18.66 |
| 6E+05 | 1E+06 | GABP01025130 | TSLPAGSASSPLTSLR | Phospho (ST)-S10 | Hypothetical protein | 32.38 |
| 6E+05 | 1E+06 | GABP01038695 | TLPASLAAPVHR | Phospho (ST)-S5 | Fatty acid desaturase | 31.34 |
| 6E+05 | 3E+05 | JO725591 | LGP MYGERSDDEGSGAEVK | Phospho (ST)-S9 Phospho (ST)-S14 | Unknown | 56.81 |
| 6E+05 | 8E+05 | GABP01061001 | GKSTESLGTDEEQAK | Phospho (ST)-S3 | 2-og-Fe Oxygenase superfamily | 92.68 |
| 6E+05 | 1E+06 | GABP01035906 | RSPLPPGPPLPPGSR | Phospho (ST)-S2 | Nucleic acid binding OB fold | 35.45 |
| 6E+05 | 3E+06 | GABP01021035 | YGLDLPSKVGK | Phospho (ST)-S7 | nischarin-like | 22.4 |
| 6E+05 | 5E+05 | GABP01107690 | TALTTVAPPAALLQVR | Phospho (ST)-T4 | Unknown | 17.48 |
| 5E+05 | 4E+05 | GABP01041893 | GESQVPTQAGDRPDSEAADETAK | Phospho (ST)-S15 | Unknown | 61.07 |
| 5E+05 | 4E+05 | GABP01038472 | AGSEEAGAAEPASNVGK | Phospho (ST)-S3 | Unknown | 101.93 |
| 5E+05 | 4E+05 | JO722680 | LPGGAMPAPMTQHMSYIQEKMNPI | Phospho (ST)-T10-T12-Y17 | cAMP-dependent protein kinase regulatory subunit | 15.57 |
| 5E+05 | 7E+05 | GABP01093776 | VFYHLEEYDDAR | Phospho (Y)-Y8 | Proteasome non atp ase subunit 1 | 33.88 |
| 5E+05 | 3E+05 | JO699741 | GLELDDVILRTPHQLTR | Phospho (ST)-T11 Deamidated [474]-Q14 | predicted protein | 23.46 |
| 5E+05 | 6E+05 | JO719542 | VGNSDLGEMAGGNSPDR | Phospho (ST)-S14 | Splicing factor, arginine/serine-rich | 64.46 |
| 5E+05 | 3E+05 | JO734709 | GLAAVLGSVALAVGTTGPR SAR | Phospho (ST)-S8-T15-T16-S20 | predicted protein | 16.37 |
| 5E+05 | 3E+05 | JO717527 | GLAKDEESEEEQEER | Phospho (ST)-S8 | pre-mRNA-processing factor 40 homolog A-like | 105.34 |
| 5E+05 | 6E+05 | JO706783 | YGEDSGDEILR | Phospho (ST)-S5 | Sas10/Utp3/C1D family | 37.52 |
| 5E+05 | 1E+04 | JO719969 | AADDDDDDDSDDEPAPK | Phospho (ST)-S10 | RNA recognition motif | 88.12 |
| 5E+05 | 1E+04 | GABP01078939 | ESIVQDLVPAKGDDSPGR | Deamidated [474]-Q5 Phospho (ST)-S15 | family protein | 17.14 |
| 5E+05 | 2E+04 | JO705826 | LSLQPVFVSLAR | Deamidated [474]-Q4 Phospho (ST)-S9 | Unknown | 38.88 |
| 5E+05 | 2E+05 | JO705182 | MEDEERSPTGPPPSK | Phospho (ST)-S7 | BTB/POZ domain | 42.84 |
| 5E+05 | 2E+05 | GABP01084598 | NLTCKLVMVLSNGR | Phospho (ST)-T3 Oxidation (M)-M8 Deamidated [474]-N12 | CSD protein | 28.31 |
| 5E+05 | 1E+05 | JO728952 | PQLSPTCLPFSAGPAQVAMPAK | Phospho (ST)-S11 Deamidated [474]-Q13 Oxidation (M)-M21 | Unknown | 18.41 |
| 5E+05 | 2E+05 | GABP01034927 | RAGDVLESIAHV GADFQELWQGA AA | Phospho (ST)-S8 | eukaryotic translation initiation factor 5B | 26.81 |
| 5E+05 | 2E+05 | GABP01025983 | HPRPAAGR | Phospho (ST)-S1 | Glycosyltransferase | 38.26 |
| 5E+05 | 2E+06 | JO706453 | SLCCIQSSNDPWNLP LK | Phospho (ST)-S7 | Unknown | 105.92 |
| 5E+05 | 3E+05 | GABP01040472 | GPVNPGSPGGWNQGGGGSGGFHR | Phospho (Y)-Y2 Deamidated [474]-N5 Carboxymethyl (C)-C11 | fibrinogen A-alpha chain | 17.01 |
| 5E+05 | 8E+05 | GABP01092376 | AY AQNRLVVVCK | Deamidated [474]-N3 Phospho (ST)-S7 | Hypothetical protein | 20.88 |
| 5E+05 | 5E+05 | GABP01081887 | TSNGTTSVALRTPK | Phospho (ST)-S10 Phospho (ST)-S15 | Cell division cycle | 87.56 |
| 5E+05 | 4E+05 | GABP01040479 | LSFKDPDANSDDGNSDAEEV PKPR | Phospho (ST)-S10 Phospho (ST)-S13 | XAP5 protein related | 34.29 |
| 5E+05 | 4E+05 | GABP01040479 | VASSMMDPESPRSPGK | | Unknown | |

| | | | | | | |
|-------|-------|--------------|---------------------------|---|---|--------|
| 5E+05 | 3E+05 | GABP01053847 | VTDDSPAAPPQK | Phospho (ST)-S5 | Amino acid transporter like | 26.47 |
| 5E+05 | 4E+05 | GABP01008094 | MISDVDDGSGTIGYEEFLK | Phospho (ST)-S10 | caltractin | 41.61 |
| 5E+05 | 8E+05 | JO725087 | STWEPQGIASAGWK | Phospho (ST)-S10 | apicomplexan-conserved protein | 51.44 |
| 4E+05 | 4E+05 | GABP01058507 | TTSSEIVQAPKPSLPGAGSR | Phospho (ST)-S3 | Unknown | 39.74 |
| 4E+05 | 5E+05 | GABP01024074 | AVSMGDVSVQSTTPVEAIHSK | Phospho (ST)-S3 | Unknown | 59.26 |
| 4E+05 | 3E+05 | GABP01019831 | FDEIEDSDDEKTQEK | Phospho (ST)-S7 | TPR domain protein | 37.03 |
| 4E+05 | 3E+05 | GABP01017850 | RICLWILSMFVLI | Carboxymethyl (C)-C3 Phospho (ST)-S8 | alpha-tubulin, partial | 16.07 |
| 4E+05 | 8E+05 | JO707225 | AFVLSFTQLAGA | Phospho (ST)-S5 Deamidated [474]-Q8 | Hsp90 | 15.98 |
| 4E+05 | 4E+05 | GABP01087383 | MTKPSLTAGPAVLR | Oxidation (M)-M1 Phospho (ST)-S5 | Unconventional myosin truncated bHLH transcription factor | 34.88 |
| 4E+05 | 5E+05 | JO733358 | AASASELLEK | Phospho (ST)-S3 | Unknown | 40.22 |
| 4E+05 | 4E+05 | GABP01079143 | GGFRMATIGGSR | Oxidation (M)-M5 Phospho (ST)-T7 | Unknown | 29.78 |
| 4E+05 | 3E+05 | GABP01107773 | SPAVVTTTGGSPITVMR | Phospho (ST)-S1 Phospho (ST)-S11 | Unknown | 46.77 |
| 4E+05 | 4E+05 | JO714961 | AGRAAPTGTGSHAAHAAASK | Phospho (ST)-T8 | Unknown | 24.19 |
| 4E+05 | 7E+05 | GABP01026436 | AVEEDEEEESDDECEIPESFKKPEA | Phospho (ST)-S10 | cAMP-dependent protein kinase regulatory subunit | 127.85 |
| 4E+05 | 3E+05 | GABP01028776 | AVTEPVPTQQASPDASPTK | Phospho (ST)-S12 Phospho (ST)-S16 | Arf-GAP like | 41.18 |
| 4E+05 | 5E+05 | GABP01025294 | AAVQSTSALAQCPSLR | Phospho (ST)-S14 | RNA pol II associated protein, partial | 18.26 |
| 4E+05 | 8E+05 | JO706757 | SRTSTSRPPR | Phospho (ST)-S6 | Unknown | 25.16 |
| 4E+05 | 1E+05 | JO707688 | AEQVELQSPAR | Phospho (ST)-S8 | cell division protein FtsY | 30.39 |
| 4E+05 | 1E+04 | GABP01091546 | ATTVVAADRPTR | Phospho (ST)-T3 | carbohydrate binding protein | 19.79 |
| 4E+05 | 1E+05 | GABP01052175 | GAAESSRVSLPAPAVAR | Phospho (ST)-S9 | dynactin | 24.74 |
| 4E+05 | 1E+04 | GABP01009202 | GLEPKNGFFVVQSSAPVPEQAK | Phospho (ST)-S14 | Unknown | 16.24 |
| 4E+05 | 1E+05 | JO719969 | KAAVEDSDDSDDEPAPK | Phospho (ST)-S7 Phospho (ST)-S11 | RNA recognition motif family protein | 71.13 |
| 4E+05 | 1E+05 | JO717160 | LATYKVPQELEAVPELPR | Phospho (ST)-T3 Phospho (Y)-Y4 | AMP-dependent synthetase and ligase | 18.68 |
| 4E+05 | 1E+04 | GABP01104632 | LQTRAPVTR | Deamidated [474]-Q2 Phospho (ST)-T8 | t complex protein 1 subunit gamma | 27.76 |
| 4E+05 | 1E+04 | GABP01021221 | NSSQCAASSVRVSLQTSR | Phospho (ST)-S8 | phosphoribosylformimino-5-aminoimidazole carboxamide ribonucleotide isomerase | 15.17 |
| 4E+05 | 1E+04 | GABP01102825 | QSLVSGTSGARPKR | Deamidated [474]-Q1 Phospho (ST)-T7 | Zn finger TFIIIS type | 20.82 |
| 4E+05 | 1E+05 | GABP01055853 | RITLWPR | Phospho (ST)-T3 | SH3 domain | 36.76 |
| 4E+05 | 2E+05 | GABP01047795 | RPPSPEGGGAAEGPDDAGSPHAGNR | Phospho (ST)-S4 Phospho (ST)-S19 Deamidated [474]-N24 | Unknown | 80.19 |
| 4E+05 | 1E+05 | JO735966 | VAAAGAEESPAK | Phospho (ST)-S10 | hypothetical protein | 61.49 |
| 4E+05 | 2E+05 | GABP01061419 | VDALDEATVQVLELSAGPLEASGLR | Phospho (ST)-S15 | Unknown | 26.64 |
| 4E+05 | 1E+04 | GABP01107499 | YEVNVPVLLHR | Phospho (Y)-Y1 Deamidated [474]-N4 | K voltage gated channel | 37.99 |
| 4E+05 | 3E+05 | GABP01076421 | AEAAAGAAAAAASALQ | Phospho (ST)-S13 | Unknown | 26.39 |
| 4E+05 | 3E+05 | GABP01025693 | LGASLALFDFAHMGASLSLR | Phospho (ST)-S4 Oxidation (M)-M13 | Unknown | 16.76 |
| 4E+05 | 4E+05 | GABP01024922 | AGDDGASDREADAGSEEEAEETGR | Phospho (ST)-S7 Phospho (ST)-S15 | pinin/sdk/mema domain containing protein | 70.34 |
| 4E+05 | 3E+05 | GABP01086521 | GADAPAVAAASPVVLPQGR | Phospho (ST)-S11 | etc interacting domain 3 | 48.18 |
| 4E+05 | 2E+05 | GABP01020598 | WGPSARAAGPRPSPGR | Phospho (ST)-S4 | PAP/25A associated domain- | 23.24 |

| | | | | | | |
|-------|-------|--------------|-----------------------------------|--|---|-------|
| | | | | | containing protein | |
| 4E+05 | 3E+05 | GABP01101024 | VRSATILTPR | Phospho (ST)-T5 | Serine protease family | 21.46 |
| 4E+05 | 5E+05 | JO743481 | WPGLLVLAQSLTPATSPAR | Phospho (ST)-S10 Phospho (ST)-T15 | saccharopine dehydrogenase | 32.08 |
| | | | QGEASALLTMSNLQANSQAPTDLR | Deamidated [474]-Q1 Phospho (ST)-S11 Phospho (ST)-S17 | TPR repeat protein | 15.95 |
| 4E+05 | 1E+06 | GABP01050228 | SAR | Phospho (ST)-T7 | SRP 68kDa protein isoform 6 similar to protein phosphatase | 71.68 |
| 4E+05 | 3E+05 | GABP01029968 | AYDDADTDDEDAKPR | Phospho (ST)-S10 | IG variant isoform 5 | 15.57 |
| 3E+05 | 4E+05 | GABP01007184 | RGGPQCQWASLWALR | Phospho (ST)-S6 Phospho (ST)-S7 | Unknown | 42.84 |
| 3E+05 | 3E+05 | GABP01046617 | EEEDPSSEEEEEAKEEGAAEK | Phospho (ST)-S4 | histone deacetylase like | 93.02 |
| 3E+05 | 2E+05 | GABP01012557 | AKGSDGEDEDEEPEPK | Phospho (ST)-S11 | Unknown | 30.66 |
| 3E+05 | 3E+05 | GABP01104468 | GLLPSSGEGDSPVDSAR | Phospho (ST)-T9 | Unknown | 44.44 |
| 3E+05 | 5E+05 | JO746453 | VAEPTGTDTPR | Oxidation (M)-M4 Oxidation (M)-M5 Phospho (ST)-S16 | Spliceosomal complex protein | 20.29 |
| 3E+05 | 3E+05 | GABP01069110 | LQVMMIVWSASHDLLSASIRAWK | Phospho (ST)-S7 | Unknown | 38.14 |
| 3E+05 | 5E+05 | GABP01008812 | IAIIFTSIVIR | Phospho (ST)-S3 | alpha-glucan water dikinase] | 42.73 |
| 3E+05 | 3E+05 | GABP01018605 | LPSWTGLFR | Phospho (ST)-S14 | Unknown | 17.03 |
| 3E+05 | 2E+05 | GABP01071950 | ALALALPPPGPASPVPRTGSGR | Phospho (ST)-S9 | C2-domain containing protein Methionine s-adenosyltransferase | 23.71 |
| 3E+05 | 2E+05 | GABP01028630 | KPAPPVGRSLR | Phospho (ST)-S5 | | 64.35 |
| 3E+05 | 3E+06 | GABP01048097 | INKQSPDIAGGVHVVGK | Deamidated [474]-N9 Phospho (ST)-S16 Oxidation (M)-M21 | Modular polyketide synthase | 15.75 |
| 3E+05 | 2E+05 | GABP01044986 | APTSMGSPNAVVPVPCSSAILMSLGRT LASR | Phospho (ST)-S8 Phospho (ST)-S10 | Structural maintenance of chromosomes4 | 33.79 |
| 3E+05 | 2E+05 | GABP01028183 | TDSPDAESESDEEPPKK | Phospho (ST)-S9 | Unknown | 55.36 |
| 3E+05 | 4E+05 | GABP01000248 | EEEEDEGASDEDVEQK | Phospho (ST)-S1 Phospho (ST)-S4 | delta-12 oleate desaturase | 22.17 |
| 3E+05 | 2E+05 | GABP01107575 | SFGSGTNLADLASQSR | Phospho (ST)-S5 | Leo1p | 57.57 |
| 3E+05 | 4E+05 | GABP01007088 | DLFGSEDEGPEIDER | Phospho (ST)-T8 | 5'-3' exoribonuclease | 16.97 |
| 3E+05 | 4E+05 | GABP01054309 | CLRPSATTSMTPPMTR | Phospho (ST)-S8 Phospho (ST)-S14 | Unknown | 25.82 |
| 3E+05 | 2E+05 | JO743725 | TVVHQRPSSPPALSPVGSAAVPTPAD PAALAK | Phospho (ST)-S16 | hypothetical protein | 64.4 |
| 3E+05 | 3E+05 | GABP01018625 | VQTMLMQPVLMTSPGSPGK | Phospho (ST)-S18 | Dynein heavy chain like | 47.86 |
| 3E+05 | 2E+05 | GABP01110361 | LVLGSPADLEIREAGLSR | Phospho (ST)-S12 | cGMP-dependent protein kinase | 53.32 |
| 3E+05 | 3E+05 | GABP01018703 | FTTGPEAEFGSPEGSEEAR | Phospho (ST)-S11 | Unknown | 93.68 |
| 3E+05 | 1E+04 | JO716142 | AAADDDSDDDSDDEPAPK | Phospho (ST)-T6 | Zf-CCCH type | 38.17 |
| 3E+05 | 1E+04 | GABP01059061 | AALRLTGSQVQR | Phospho (ST)-S3 Phospho (ST)-S7 | RNA recognition motif family protein [| 22.42 |
| 3E+05 | 9E+04 | JO719969 | ADSDEESEEEPPQKK | Phospho (ST)-T16 | MAPK2, putative uncharacterized protein | 48.2 |
| 3E+05 | 1E+04 | JO739971 | AIGAAEQPHLQALPNTPR | Phospho (ST)-T10 | LOC100273704 | 23.62 |
| 3E+05 | 7E+04 | GABP01021399 | APAAGPPGATRPGR | Phospho (ST)-T7 | Pyruvate kinase | 56.5 |
| 3E+05 | 1E+04 | GABP01071379 | GVLPVLTQSFVGTDSVIK | Phospho (ST)-S7 Deamidated [474]-Q10 | Major Facilitator Superfamily | 27.45 |
| 3E+05 | 1E+05 | JO709869 | KPKPLKSDAQR | Phospho (ST)-T10 | EF-1 alpha | 23.22 |
| 3E+05 | 1E+04 | GABP01040147 | LTIPSLLLGTR | Deamidated [474]-N2 Phospho (ST)-S7 Phospho (ST)-S17 | conserved hypothetical protein | 19.85 |
| 3E+05 | 1E+04 | JO714264 | NNELIFSVNPPPLGELSAMKGK | Phospho (ST)-T8 | hypothetical protein | 21.09 |
| 3E+05 | 1E+04 | JO712236 | RETSGMVTALR | Deamidated [474]- | Phosphoglycerate kinase | 16.9 |
| 3E+05 | 1E+04 | GABP01059807 | SVLGANAILAVSMAVCRAAAASEM | | | |

| | | | | | | |
|-------|-------|--------------|-----------------------------|---|--|--------|
| | | | PLYQYIAK | N6 Phospho (ST)-S12 Oxidation (M)-M25 | | |
| | | | | Phospho (ST)-S3 Carboxymethyl (C)-C5 Phospho (ST)-T9 | ML superfamily protein (Cysteine protease like) | 31.3 |
| 3E+05 | 1E+05 | GABP01008016 | TGSVCEDSTITLPLGLGK | Carboxymethyl (C)-C5 Phospho (ST)-T6 | Unknown | 17.44 |
| 3E+05 | 1E+04 | GABP01084233 | TSPCCTVARSTACTPAEAVR | Phospho (ST)-T12 | t complex protein 1 subunit alpha | 29.79 |
| 3E+05 | 1E+04 | GABP01084658 | VCTSRITPGPSTPTSNGPSQSSR | Deamidated [474]-Q9 Phospho (ST)-S16 Phospho (ST)-T22 | Unknown | 27.98 |
| 3E+05 | 4E+04 | GABP01068279 | WTKATPLAQGGDEACSQLPASTWG K | Phospho (ST)-T5 | step II splicing factor SLU7, putative | 17.49 |
| 3E+05 | 5E+05 | GABP01014465 | SLRPTASQRCAAR | Phospho (ST)-S10 | Chromosome segregation ATPases | 23.26 |
| 3E+05 | 7E+05 | JO711605 | SAVIGPRTASEK | Phospho (ST)-S3 | Nucleolar protein Nop5, putative | 102.47 |
| 3E+05 | 2E+05 | GABP01062475 | KESEAEAAAEAEAPPPEK | Phospho (ST)-S20 | eukaryotic translation initiation factor 3 p42 subunit | 21.58 |
| 3E+05 | 2E+05 | JO721439 | SDASCCCMSPPLPPSLPPWLP R | Phospho (ST)-T17 Phospho (ST)-S21 Carboxymethyl (C)-C23 | Chalcone and Stilbene domain containing | 16.37 |
| 3E+05 | 2E+05 | GABP01078304 | GSGRSTRSRPAASLGGTIAGSFCWT | Phospho (ST)-S14 | malate:quinone oxidoreductase | 20.58 |
| 3E+05 | 9E+05 | JO762256 | LSTGSEDLLNQRPMSK | Phospho (ST)-T3 | SET domain containing protein | 32.11 |
| 3E+05 | 2E+05 | GABP01027848 | SCTVGCCPHPTLPPPPRPR | Phospho (ST)-S3 | PRP38 family | 29.96 |
| 3E+05 | 4E+05 | JO699504 | AISEDRDADAR | Phospho (ST)-S14 | Calcium dependent protein kinase domain | 62.04 |
| 3E+05 | 4E+05 | GABP01049570 | APDDTSMFDRYPESTEGSAPSISQAD | Phospho (ST)-S8 Phospho (ST)-S12 | Cold Shock domain containing protein | 18.94 |
| 3E+05 | 4E+05 | GABP01049570 | QEHFEFGFK | Phospho (ST)-S7 Phospho (ST)-T10 | Unknown | 47.82 |
| 3E+05 | 3E+05 | GABP01054190 | LAPHSAAPEPSRPRSPEAER | Oxidation (M)-M10 Deamidated [474]-Q19 Phospho (ST)-T28 | Unknown | 17.12 |
| 3E+05 | 2E+05 | GABP01041893 | AAAPAESAPTDEQAGR | Phospho (ST)-T17 | snRNP GAR1 protein | 54.63 |
| 3E+05 | 4E+05 | GABP01113699 | SNACPPPALMAASVSSPSQALTALQ | Phospho (ST)-T5 | Unknown | 41.22 |
| 3E+05 | 2E+05 | GABP01036833 | VTMLGVMWR | Phospho (ST)-S5 | Unknown | 65.72 |
| 3E+05 | 2E+05 | GABP01013748 | AAQEPAAGEDQGEAAPTTPPAAAGS K | Phospho (ST)-T9 | thrombospondin type 1 repeat containing protein | 18.6 |
| 3E+05 | 2E+05 | GABP01013748 | APGGTPPAELEAAPVASQEV R | Deamidated [474]-N3 Phospho (ST)-T5 | Cell division cycle | 39.37 |
| 3E+05 | 3E+05 | GABP01040066 | HMQESDEESPTGLEK | Phospho (ST)-T4 | hypothetical protein | 15.77 |
| 3E+05 | 3E+05 | GABP01011686 | QPPATIPHTR | Phospho (ST)-S3 | glutamate decarboxylase | 24.01 |
| 3E+05 | 4E+05 | GABP01092376 | TSNGTTSVALRTPK | Phospho (ST)-T7 | thrombospondin type 1 repeat containing protein | 20.65 |
| 3E+05 | 3E+05 | GABP01112826 | FVSTAPKGGGVLSR | Phospho (ST)-T2 Oxidation (M)-M19 Phospho (ST)-S20 | Unknown | 16.81 |
| 3E+05 | 3E+05 | JO721849 | QVSVPPYYSDLTR | Phospho (ST)-S5 | seven thrombospondin repeats (type 1 and type 1-like) | 38.86 |
| 3E+05 | 3E+05 | GABP01058661 | LLLTTTACVCLSGLQHACK | Phospho (ST)-S3 | Unknown | 88.71 |
| 3E+05 | 2E+05 | JO721667 | STGRPAQPAPAPSATALKMSAPR | Phospho (ST)-S3 | Unknown | 49.56 |
| 3E+05 | 2E+05 | JO711723 | AGEESPITVAK | Phospho (ST)-S1 | alpha-glucan water dikinase | 73.39 |
| 3E+05 | 3E+05 | GABP01063423 | SDSPDALGDFNSGQQGGGR | Deamidated [474]-N9 | RRM domain containing | 89.94 |
| 3E+05 | 5E+05 | GABP01027440 | GGSDDGLPSGGSDVEAYISR | Phospho (ST)-S13 | | |

| DGK | | | | protein | | | |
|-------|-------|--------------|---|---|--|--|--------|
| 3E+05 | 2E+05 | GABP01086264 | ITEEINMLKSLK | Phospho (ST)-T2 Deamidated [474]-N6 Oxidation (M)-M7 | Protein kinase catalytic domain | | 23.3 |
| 3E+05 | 3E+05 | GABP01100888 | VK | Phospho (ST)-S24 | Arf GAP like | | 65.55 |
| 3E+05 | 3E+05 | GABP01019516 | TPSTTATEESDGEPPDDLHLTLAP | Phospho (ST)-S10 | hypothetical protein | | 49.94 |
| 2E+05 | 2E+05 | JO714359 | AEPAQPK | Phospho (ST)-S4 | hnRNP like | | 21.54 |
| 2E+05 | 2E+05 | GABP01081001 | EQCQCGDNMTLMVVDLQAGGAGCPAASGGLQA ASPGTASVNGAAQ | Phospho (ST)-S27 Deamidated [474]-Q31 | Protein phosphatase | | 16.94 |
| 2E+05 | 2E+05 | GABP01029502 | LLFAGPAVTRTLAGESLAPSSSTSSPSA | Phospho (ST)-T9 Phospho (ST)-T11 | Coagulation factor 5/8 C terminal domain containing protein | | 44.16 |
| 2E+05 | 2E+05 | JO749092 | STPSSAR | Phospho (ST)-S4 | acetyl-CoA carboxylase | | 16.57 |
| 2E+05 | 1E+05 | GABP01096459 | ISSSGGALCPSAAWLAPR | Phospho (ST)-S9 | Kinesin like | | 16.48 |
| 2E+05 | 3E+05 | GABP01024922 | SALCRPARSSGAGPR | Phospho (ST)-S17 | pinin/sdk/mema domain containing protein | | 134.06 |
| 2E+05 | 7E+05 | GABP01031190 | GAAGEAAEAADAEAAASPGEEGK | Phospho (ST)-S5 Deamidated [474]-Q9 Phospho (ST)-S23 | DEAD box RNA/DNA helicase | | 16.8 |
| 2E+05 | 4E+05 | JO761509 | GAPGSVTSQHRATIAAMELNCSR | Deamidated [474]-Q3 Phospho (ST)-S9 Oxidation (M)-M10 | serine/threonine protein kinase related protein [DNAj domain, possible | | 35.9 |
| 2E+05 | 3E+05 | JO754053 | LDQITLPVSMRLR | Phospho (ST)-S10 | transmembrane domain | | 46.51 |
| 2E+05 | 4E+05 | JO707869 | ASGFTEDNNSDDEQEYQPDEPAEVV | Phospho (ST)-T9 | inosine-5'-monophosphate dehydrogenase 2-like | | 35.59 |
| 2E+05 | 1E+05 | JO709540 | RAGAAPRPTR | Phospho (ST)-T21 | Pumilio-family RNA binding domain | | 15.3 |
| 2E+05 | 3E+05 | GABP01036474 | GICGGNMPYIEGSQAKSPAGTVLTLR | Phospho (ST)-S19 | Unknown | | 57.73 |
| 2E+05 | 1E+05 | GABP01047267 | SAAPVETPADGAVTAAASPR | Phospho (ST)-S12 | Unknown | | 46.14 |
| 2E+05 | 2E+05 | GABP01084988 | AQAPKPPGSMPTPAGEVAAAK | Phospho (ST)-S7 | CDC 37 like protein | | 104.77 |
| 2E+05 | 2E+05 | GABP01061067 | WDNIELSDDES DLHPNIDK | Phospho (ST)-T19 | Unknown | | 26.62 |
| 2E+05 | 2E+05 | GABP01061823 | DAPPIGPRHAPARPEVTGTGSVQA | Phospho (ST)-T20 | peptidase m16 domain containing | | 35.25 |
| 2E+05 | 5E+05 | GABP01049570 | SLACAIPGEVAREASSVAGTR | Phospho (ST)-T3 | Unknown | | 83.25 |
| 2E+05 | 2E+05 | GABP01094006 | TWTLCGTPEYLAPEIIQSK | Phospho (ST)-S10 | Unknown | | 92.63 |
| 2E+05 | 4E+05 | GABP01041205 | VSAQHDAAGSGDDAGEAR | Phospho (ST)-S5 | Unknown | | 45.58 |
| 2E+05 | 2E+05 | GABP01025130 | WADCSDEDEEDER | Phospho (ST)-S17 | Hypothetical protein | | 18.6 |
| 2E+05 | 4E+05 | JO707767 | WADATPTMAFTPV AEDSR | Oxidation (M)-M1 Phospho (ST)-T4-S11 | tetratricopeptide repeat-containing protein | | 22.82 |
| 2E+05 | 7E+05 | GABP01080861 | MIGTPCQRALSTAAAVAAQPR | Phospho (ST)-S3 | Myotubularin related EF hand protein (Calcium ion binding) | | 38.84 |
| 2E+05 | 1E+04 | GABP01025015 | KASQEDAWKEPK | Phospho (ST)-T6 | hypothetical protein | | 22.61 |
| 2E+05 | 1E+04 | JO717527 | AAPNLTIPLSAGPSKSK | Phospho (ST)-S4 | hypothetical protein | | 42.35 |
| 2E+05 | 4E+04 | GABP01066887 | DEESEEQEER | Phospho (ST)-S11 | Armadillo type fold | | 55.82 |
| 2E+05 | 1E+05 | JO743508 | EGEGAQGAEASPAASK | Phospho (ST)-T5 Phospho (ST)-S7 | heat shock protein 90 1 | | 40.88 |
| 2E+05 | 1E+04 | GABP01046561 | EKEVTDSEDEEEK | Phospho (ST)-S4 Phospho (ST)-S5 | TRP repeat protein | | 65.12 |
| 2E+05 | 7E+04 | GABP01045542 | GGNSSADEHESSTVAR | Phospho (ST)-S19 Deamidated [474]-Q26 Oxidation (M)-M29 | 60S ribosomal protein | | 16.34 |
| 2E+05 | 1E+04 | JO709540 | HVQTS MAYCARQTL MQCTSGAISP | Phospho (ST)-S10 Phospho (ST)-S21 | hypothetical protein | | 63.52 |
| 2E+05 | 1E+04 | JO709548 | QMGAMSR | Deamidated [474]- | RRM domain protein | | 18.59 |
| 2E+05 | 1E+04 | JO709548 | KGGAGQGEASSPLAAEQEPLSPGAA | | | | |
| 2E+05 | 1E+04 | JO709548 | AEDAGAGR | | | | |
| 2E+05 | 1E+04 | JO709548 | QTVPELLSHMPQAVHTSQAGIAANA | | | | |

| | | | | | | |
|-------|-------|--------------|--|---|---|--------|
| | | | LAQAR | Q1 Phospho (ST)- T2 Phospho (ST)-S17 | | |
| 2E+05 | 1E+04 | JO707688 | RGASPPSPASCPQGGSPGAR | Phospho (ST)-S4-S7-S16 | cell division protein Fts | 15.42 |
| | | | RILPMFSSSQWGFVGCSEMVFCFS | Oxidation (M)-M5 Phospho (ST)-S14 Carboxymethyl (C)-C24 | small COPII coat GTPase SAR1 | |
| 2E+05 | 6E+04 | GABP01019272 | TCR | Phospho (ST)- T10 Deamidated [474]-N13 | | 17.09 |
| 2E+05 | 5E+04 | GABP01031902 | SSAAPQSSPTAANGAAK | Phospho (ST)-S3 | Unknown | 56.05 |
| 2E+05 | 1E+04 | GABP01096787 | TASGEIDRGEPGNAPPVEEIEK | | RRM and KH domain protein similar to putative chromatin structure regulator | 55.69 |
| 2E+05 | 1E+04 | JO757628 | TQAAESREDLAAELFGDDVDADHAPHER | Phospho (ST)-S6 | | 53.01 |
| 2E+05 | 1E+04 | JO743752 | VDAIVESLK | Phospho (ST)-S7 | 50S ribosomal protein L12 glycerol-3-phosphate | 21.85 |
| 2E+05 | 1E+06 | JO705455 | LSEVFNESKENIK | Phospho (ST)-S8 | dehydrogenase | 32.39 |
| | | | GFAPQTPAGLGQMPATPAPPHVPATP | | similar to putative chromatin structure regulator | |
| 2E+05 | 2E+05 | JO757628 | AGVAPMSPAR | Phospho (ST)-S33 | | 59.97 |
| | | | | Deamidated [474]- Q2 Deamidated [474]- Q3 Phospho (ST)-T7 | G8 domain, Right handed beta helix region | |
| 2E+05 | 3E+05 | JO704970 | RQQSWFTLPR | Phospho (ST)-S19 Phospho (ST)-S20 | | 16.15 |
| 2E+05 | 2E+05 | JO719542 | VPAHGGAPQVDKVEEMAFSSEED | | Immediate-early protein | 39.98 |
| 2E+05 | 1E+05 | GABP01108500 | CSGSARGAPPGPATR | Phospho (ST)-T15 | Hypothetical | 20.16 |
| 2E+05 | 3E+06 | JO706453 | GPANPGSPGAIGGWR | Phospho (ST)-S7 | fibrinogen A-alpha chain | 65.6 |
| 2E+05 | 7E+05 | GABP01111212 | TPAAGARPRSPLR | Phospho (ST)-T1 | Unknown | 22.46 |
| | | | | Deamidated [474]- N9 Phospho (ST)-T12-S27- T34-Y41 | TypeI polyketide synthase | |
| 2E+05 | 2E+05 | GABP01055210 | MIGLEPTNCATHAAVCMVHPVQLCSLRMCGMTSSQLLAYVAVIMPR | Phospho (ST)-S5-S11-S12- T13-T15-S16-S17-S18- S19-T21 | | 15.4 |
| 2E+05 | 9E+04 | GABP01041886 | AMSSSCAACGSSTATSSSGTSTR | | Unknown | 18.82 |
| 2E+05 | 3E+05 | GABP01000248 | YAPINVLDSDDEADAPLPPPPK | Phospho (ST)-S9 | Unknown | 104.21 |
| 2E+05 | 2E+05 | GABP01086127 | NSAGPLAVRPTHVQAACTMR | | Zeaxanthin chloroplast precursor | |
| 2E+05 | 5E+05 | GABP01076806 | LPTALSR | Phospho (ST)-T18 | | 16.22 |
| | | | | Phospho (ST)-T3 | Armadillo like | 20.96 |
| 2E+05 | 2E+05 | GABP01041676 | VPVLSVCALSK | Carboxymethyl (C)- C7 Phospho (ST)-S10 | | |
| 2E+05 | 4E+05 | GABP01004920 | SLSFDSPSGPPSK | tyrosine aminotransferase | | 19.61 |
| 2E+05 | 1E+05 | GABP01092058 | STSRPPAQSSGGTSPVAER | Phospho (ST)-S3 | Unknown | 35.99 |
| | | | | Phospho (ST)-S10 | Unknown | 29.68 |
| 2E+05 | 2E+05 | GABP01038926 | RGADGGDGASGESEGEVER | Phospho (ST)-S10 Phospho (ST)-S13 | Formin binding domain containing | 48.8 |
| | | | GYGPEEDLHGSFPDWSVEPEALWEF | | | |
| 2E+05 | 3E+05 | JO750329 | ADDAESR | Phospho (ST)-S15 | carbonic anhydrase 2 | 98.46 |
| | | | | Phospho (ST)-S2 Phospho (ST)-S3 Deamidated [474]- Q14 | | |
| 2E+05 | 1E+06 | JO739000 | SSSAAPLAAAASAQHHPCTRGK | Deamidated [474]- N1 Phospho (ST)- S22 Phospho (ST)-S25 | Unknown | 24.13 |
| 2E+05 | 3E+05 | GABP01106531 | NGSPSLLTSSTGLLCGSASSGSVRSLR | | | 24.31 |
| 2E+05 | 8E+04 | JO719720 | SLVEEGAPSVFVLADK | Phospho (ST)-S1 | Unknown | 15.23 |
| | | | | | SET domain containing protein | |
| 1E+05 | 4E+05 | GABP01086031 | TTTVSAQASTKAAASMALR | Phospho (ST)-S5 | | 27.93 |
| 1E+05 | 1E+05 | JO731992 | IVEVPTVCTQEVVKAVPK | Phospho (ST)- T9 Deamidated [474]-Q10 | | |
| 1E+05 | 2E+05 | JO695062 | GVLPLVTQSFVGTDSVIK | | ALVEOLIN1 | 23.49 |
| | | | | Phospho (ST)-T7 | Pyruvate kinase | 76.57 |
| 1E+05 | 3E+05 | GABP01114472 | FHSVQDETFLGDTK | | Endoplasmic precursor(HSP 90 like) | 69.03 |
| 1E+05 | 3E+05 | GABP01012985 | DPTSPTATAGSVEYNILTGK | Phospho (ST)-S3 | | |
| | | | | Phospho (ST)-T3 | calcium-binding protein | 27.94 |
| 1E+05 | 8E+04 | GABP01046365 | LGELSAMIEATK | Phospho (ST)-S5 Oxidation | Unknown | 17.05 |

| | | | | | | |
|-------|-------|--------------|-----------------------------|---|---|-------|
| | | | | (M)-M7 | | |
| 1E+05 | 2E+05 | GABP01038074 | GRAEEEEETPASK | Phospho (ST)-T10 | Epoxy hydrolase domain containing protein | 49.49 |
| 1E+05 | 2E+05 | GABP01052524 | ATTTAISPRLAR | Phospho (ST)-S7 | monooxygenase fad-binding | 44.15 |
| 1E+05 | 3E+05 | GABP01007088 | EIFGDISDDEEPEKVEDVILR | Phospho (ST)-S7 | Leo1p | 76.7 |
| 1E+05 | 1E+05 | GABP01079817 | VRPARPRPSR | Phospho (ST)-S9 | U | 16.34 |
| | | | VTVAEAPSLWPWASLAMPASAAAE | Phospho (ST)-S8 Oxidation (M)-M17 Phospho (ST)-T24 | Unknown | 27.27 |
| 1E+05 | 2E+05 | GABP01093834 | ARK | | | |
| 1E+05 | 3E+05 | GABP01036009 | ISAVIESVPDKSPR | Phospho (ST)-S12 | Glycosyl transferase family2 | 57.47 |
| 1E+05 | 8E+04 | GABP01085385 | AQQAEPQASPEAK | Phospho (ST)-S11 | HMG protein 1 | 72.69 |
| | | | QRPGPSAPPSAPASGRPSVVGYPGM | | | |
| 1E+05 | 3E+05 | GABP01018625 | PPKSPANR | Phospho (ST)-S30 | Unknown | 37.72 |
| | | | ICNQLIPSHLQALLMR | Carboxymethyl (C)-C2 Phospho (ST)-S8 | probable ubiquitin-specific processing protease 21 | 18.49 |
| 1E+05 | 2E+05 | GABP01017846 | HNDDEQYIWESGAGGSFTVQKDTL | | | |
| 1E+05 | 1E+05 | GABP01003893 | VHGEVK | Phospho (ST)-T23 | Hsp 90 | 92.62 |
| | | | FGATRSILALR | Phospho (ST)-S6 | Nucleolar protein Nop56, putative | 17.22 |
| 1E+05 | 5E+05 | JO758863 | | | | |
| 1E+05 | 4E+05 | JO731589 | LPTLVVALLAPAS | Phospho (ST)-S13 | Alkylated dna repair protein mitochondrial ATP synthase | 15.96 |
| | | | SSGRTTASSTPR | Phospho (ST)-S1 | F1 alpha subunit-like protein | 20.26 |
| | | | ASRPPLQLELERSAGKPAVSPLPR | Deamidated [474]-Q7 Phospho (ST)-S20 | | 17.36 |
| 1E+05 | 2E+05 | JO738704 | | | lactaldehyde reductase | |
| 1E+05 | 1E+05 | JO754990 | GRCAALLAADLLGPYEAVALR | Phospho (Y)-Y15 | Cellulose synthase | 24.53 |
| | | | AGRPSTR | Phospho (ST)-S6 | cAMP dependent protein kinase regulatory | 26.12 |
| 1E+05 | 1E+05 | GABP01101185 | | | Soluble starch synthase 1, chloroplastic/amyloplastic | 64.32 |
| | | | HISSFSALSIMAEAAEKDEAR | Phospho (ST)-S4 | | |
| | | | LSEAPQKPAVQQASESATAAADEAL | | | |
| 1E+05 | 9E+04 | JO727148 | AILR | Phospho (ST)-S2-S14-S16 | TPR domain protein | 15.45 |
| 1E+05 | 2E+05 | GABP01000748 | AESDDVAADVAFAPMETDEEVAR | Phospho (ST)-S3 | protein kinase | 91.53 |
| 1E+05 | 8E+04 | GABP01099160 | PELRSPGR | Phospho (ST)-S5 | Arginyl tRNA synthetase | 18.97 |
| 1E+05 | 2E+05 | JO710748 | NIMSFSALTEMAQPVVK | Phospho (ST)-S4 | starch synthase isoform I | 46.39 |
| | | | EESPQPSGHGDPYEGFGSLPPDEFLLP | Phospho (ST)-S3 Deamidated [474]-Q5 | Unknown | 55.75 |
| 1E+05 | 2E+05 | GABP01036280 | QEMDLR | | NADP-dependent isocitrate dehydrogenase | 16.27 |
| 1E+05 | 4E+05 | JO703826 | MSRPVSVWNFRDSLTLAMTLSCAR | Phospho (ST)-S2-S7-S13 | | |
| | | | IEDLSAQAAQASAAQQFTQQGLEAAV | Phospho (ST)-S11 Deamidated [474]-Q15 Phospho (ST)-T17 | Nascent polypeptide-associated complex subunit alpha | 43.8 |
| 1E+05 | 1E+05 | JO734675 | EAAETSAEAPK | | | |
| 1E+05 | 1E+05 | GABP01033307 | SPSRDGSPPNDWIVNMLIDR | Phospho (ST)-S1-S3-S7 | Unknown | 16.81 |
| | | | SNACPPPALMAASVSSPSQALTTALQ | Oxidation (M)-M10 Phospho (ST)-S18 Deamidated [474]-Q26 | Unknown | 29.15 |
| 1E+05 | 2E+05 | GABP01113699 | VTMLGVMWR | | | |
| 1E+05 | 5E+04 | JO737776 | AAGGEASEPGTPSK | Phospho (ST)-T11 | Unknown | 45 |
| 1E+05 | 1E+04 | GABP01055853 | AASAAAQQPASPSR | Phospho (ST)-S11 | SH3 domain | 43.9 |
| 1E+05 | 1E+04 | JO751885 | AEEEEETPASK | Phospho (ST)-T8 | Unknown | 40.52 |
| 1E+05 | 4E+04 | GABP01038927 | DEDESEDEQEPK | Phospho (ST)-S4 | WW domain (formin binding) | 56.44 |
| 1E+05 | 1E+04 | GABP01033303 | IITSSTQSVPPASRR | Phospho (ST)-S8 | RUBISCO | 15.82 |
| | | | ITTWLNSKASK | Phospho (ST)-T3 Deamidated [474]-N6 | Unknown | 17.94 |
| 1E+05 | 1E+05 | JO713775 | LCARSCACQALVSEACAAGAAFFL | Carboxymethyl (C)-C2 Phospho (ST)-S5 | 5'-3' exoribonuclease | 16.19 |
| 1E+05 | 1E+04 | GABP01054309 | AGAGVAR | Deamidated [474]-N1 Oxidation (M)-M2 Phospho (ST)-S3 | Unknown | 15.53 |

| | | | | | | |
|-------|-------|--------------|--|---|--|-------|
| 1E+05 | 1E+04 | GABP01048293 | QAEESGEENAPATK | Phospho (ST)-S5 | digestive organ expansion factor | 77.62 |
| 1E+05 | 6E+04 | GABP01097283 | QLLAGLMVWPGGTWRSASAGTPGT APTK | Phospho (ST)-S18 Phospho (ST)-T24 | Unknown | 23.05 |
| 1E+05 | 1E+04 | GABP01109842 | SLCCMPNPRIGQGILSSQR | Phospho (ST)-S1 | snoRNA binding (Ribosome biogenesis) | 22.96 |
| 1E+05 | 6E+04 | GABP01082082 | TEEAAAPSPK | Phospho (ST)-S8 | actin binding (gelsolin domain) | 36.99 |
| 1E+05 | 1E+04 | GABP01051289 | TGAGAAAGGADGEDGRASPR | Phospho (ST)-S17 | Unknown | 60.77 |
| 1E+05 | 1E+04 | GABP01113382 | TGVAALAPLYQASASTR | Deamidated [474]-Q12 Phospho (ST)-S14 | C2 calcium dependent | 38.82 |
| 1E+05 | 9E+04 | GABP01007146 | ALNIVLASGLAK | Deamidated [474]-N3 Phospho (ST)-S8 | Unknown | 21.2 |
| 1E+05 | 8E+04 | GABP01028980 | QSPQPPIPTTSHQRPPCQTRSALLSS AR | Phospho (ST)-S2 Carboxymethyl (C)-C17 | solute carrier protein35 member e4 | 18.65 |
| 9E+04 | 2E+05 | JO754337 | DLSEGELEDLEDCLDAVQRPFLLQR | Phospho (ST)-S3 | Unknown | 86.81 |
| 9E+04 | 7E+04 | GABP01073537 | RMACSFCFLNASAWAFHLCTSSIGA CALTTSSSTR | Phospho (ST)-S5-Y6-S12-T20-S31-S32-S33-T34 | Unknown | 20.26 |
| 9E+04 | 2E+05 | GABP01105472 | LQLNMAYTEPPTQNLSSDYSK | Oxidation (M)-M5 Phospho (ST)-T8 Phospho (Y)-Y19 | Enolase 1 | 17.17 |
| 9E+04 | 2E+05 | GABP01022683 | DPLLLGAGSAMGPKVSEPIA | Phospho (ST)-S16 | Monovalent cation:proton antiporter family | 46.08 |
| 9E+04 | 1E+05 | GABP01075518 | RSCSMSASCNGNSCSCGVSGNSCLY CLR | Phospho (ST)-S2-S6-S8-S13-S15-S19-S22-Y25 | Unknown | 16.33 |
| 8E+04 | 5E+04 | GABP01019692 | RHSIHSQDLGLVHTLR | Phospho (ST)-S3 Deamidated [474]-Q7 | Unknown | 16.7 |
| 8E+04 | 1E+05 | JO693843 | AAEAAEDADTDMGEETLTAATLPLE GAAR | Phospho (ST)-T10 | WGR domain of poly(ADP-ribose) polymerases | 80.5 |
| 8E+04 | 1E+04 | GABP01025704 | MGSSLSVLEFAHLSASLSLR | Oxidation (M)-M1 Phospho (ST)-S4 Phospho (ST)-S16 | Unknown | 16.62 |
| 8E+04 | 4E+04 | GABP01030858 | MQLQQITPLQPPVPSSRR | Deamidated [474]-Q5 Phospho (ST)-S16 Phospho (ST)-S17 | U2 snRNP auxiliary protein | 16.01 |
| 8E+04 | 1E+05 | GABP01061399 | LPSGLPDCTPTASEELGAWR | Phospho (ST)-T9 | Unknown | 55.81 |
| 7E+04 | 1E+04 | GABP01063933 | SSSCSNRSGAGAPR | Phospho (ST)-S8 | Unknown | 18.16 |
| 7E+04 | 7E+04 | GABP01043687 | YPVSLSTLNTSGNDVFNSSCGNYWSAVPRCAQE GLGVPSLLAGSSK | Phospho (ST)-S20 | Unknown | 17.29 |
| 7E+04 | 2E+05 | JO710259 | SRPAVASRSQLQR | Phospho (ST)-S9 Deamidated [474]-Q10 | Unknown | 22.72 |
| 7E+04 | 1E+05 | GABP01111560 | VIIHELLSR | S9 Deamidated [474]-Q10 | malonyl-CoA:ACP transacylase | 44.38 |
| 6E+04 | 7E+04 | GABP01000274 | ISVTSVTFHPAGGASLPVL | Phospho (ST)-T8 | Unknown | 21.77 |
| 6E+04 | 1E+04 | JO747619 | KPEDEASDNDGSEEAK | Phospho (ST)-S7 Phospho (ST)-S12 | Unknown | 16.4 |
| 5E+04 | 8E+04 | GABP01029969 | FQITALDEPMELTSACASWR | Deamidated [474]-Q2 Phospho (ST)-S14 | Tetkin domain containing protein | 22.34 |
| 5E+04 | 1E+04 | GABP01068403 | ASRGSAASSAGATGAGTSGTSSSS WPR | Phospho (ST)-S9-S10-T14-T18-S19-T21-S22-S23 | Unknown | 20.33 |
| 5E+04 | 1E+04 | GABP01046571 | RLPGLAGGHAAASPSR | Phospho (ST)-S16 | Unknown | 38.11 |
| 5E+04 | 1E+04 | GABP01093678 | RPPPSPPR | Phospho (ST)-S5 | 2 og-fe oxidoreductase like | 62.41 |
| 5E+04 | 1E+04 | GABP01062726 | SSRPPTTGPR | Phospho (ST)-S1 | afg1 family atpase | 16.07 |
| 4E+04 | 1E+04 | JO734365 | DTVTRGVSGQATGSEVR | Phospho (ST)-S7 Deamidated [474]-Q9 | Unknown | 45.47 |
| 4E+04 | 4E+04 | GABP01004963 | AWPSGTANPSAGTILSTASPPSK | Phospho (ST)-S4-T6-T14-T18-S20-S23 | ubiquitin carboxyl-terminal hydrolase isozyme L5 | 18.6 |
| 3E+04 | 1E+04 | GABP01024579 | NAPSGSGSAADWSGAR | Deamidated [474]-N1 Phospho (ST)-S13 | Myosin XI | 16.35 |
| 1E+04 | 7E+05 | GABP01033771 | AKPRMLPGPGY AALQETDSAR | Phospho (Y)-Y11 Deamidated [474]-Q15 | Isopropylmalate isomerase | 18.36 |

| | | | | | | |
|-------|-------|--------------|----------------------------------|---|---|-------|
| 1E+04 | 2E+06 | GABP01029401 | ALQELRPCALVQLGTAIHMYVGR | Deamidated [474]- Q12 Phospho (ST)- T15 Phospho (Y)-Y20 | Kelch domain containing | 20.77 |
| 1E+04 | 4E+05 | GABP01033538 | ALTRTQRPAIGGMQTPMR | Phospho (ST)-T3 Oxidation (M)-M18 | Unknown | 31.78 |
| 1E+04 | 3E+05 | GABP01069616 | AREEGGEAPGSPLEEHLGDEEPGAEA | Phospho (ST)-S11 | RRM-CSP-RAN BP2-OB fold | 94.89 |
| 1E+04 | 4E+05 | GABP01018625 | ASSPAQFVQPK | Phospho (ST)-S3 | Unknown | 15.02 |
| 1E+04 | 3E+04 | GABP01087772 | ASSWRPPGASSR | Phospho (ST)-S2 | Unknown | 40.19 |
| 1E+04 | 9E+04 | GABP01060877 | ASVSAEAFGDWNRK | Phospho (ST)-S4 | cAMP dependent PK regulatory subunit | 58.05 |
| 1E+04 | 9E+05 | GABP01018411 | ATAKSLGASGGIATSLCTLK | Phospho (ST)-T18 | clathrin heavy chain | 16.09 |
| 1E+04 | 2E+05 | GABP01031792 | AVADDDDDFSDFDIESDAEEATAK | Phospho (ST)-S10 Phospho (ST)-S16 | CSD protein | 34.65 |
| 1E+04 | 2E+05 | GABP01028776 | AVTEPVPTQQASPDASPTK | Phospho (ST)-S12 Phospho (ST)-T18 | ARF-GAP like | 20.12 |
| 1E+04 | 2E+06 | GABP01060639 | AWPPGSSFNPPGTSALR | Deamidated [474]- N9 Phospho (ST)-S14 | Unknown | 21.3 |
| 1E+04 | 3E+05 | GABP01043186 | AYS GDIPDRPVEGAPSEVPTSR | Phospho (ST)-S3 | RRM/CSp/OB fold non discriminatory gln-glu tRNA synthetase | 52.58 |
| 1E+04 | 2E+05 | GABP01097512 | CAPAAATAGTRAGK | Phospho (ST)-T7 | | 24.2 |
| 1E+04 | 2E+06 | GABP01030014 | CCGRRPAARVLPPTGTP | Phospho (ST)-T16 | s/t protein kinase Leucine Rich Repeat family protein | 21.2 |
| 1E+04 | 3E+05 | GABP01015966 | CLPSRSSLR | Phospho (ST)-S7 | | 27.93 |
| 1E+04 | 8E+04 | GABP01110879 | DASSLQTLISGLQMLSISGLPCGRTC | Phospho (ST)-S3 Phospho (ST)-T26 | Dual specificity phosphatase domain | 32.02 |
| 1E+04 | 7E+04 | GABP01020434 | DLTGRSHR | Phospho (ST)-S6 | Unknown | 17.92 |
| 1E+04 | 9E+04 | GABP01056264 | DYQAAAASANQAR | Phospho (ST)-S8 Deamidated [474]-Q11 | TPR repeat protein | 20.18 |
| 1E+04 | 2E+05 | JO724751 | ECVANEGGESPVKKPPDDSTDFEQI SK | Phospho (ST)-S10 | Unknown | 45.1 |
| 1E+04 | 1E+05 | JO717524 | EFLVHEEPASPSTAALAEPPAAK | Phospho (ST)-S10 | Unknown | 57.01 |
| 1E+04 | 7E+05 | GABP01032554 | EQHVQSVLINPNIA TVQTSK | Phospho (ST)-T18 Phospho (ST)-S19 | Carbamoyl phosphate synthase leucyl-tRNA synthetase, cytoplasmic isoform 4 | 18.5 |
| 1E+04 | 3E+05 | JO720290 | FLERATK | Phospho (ST)-T6 | | 15.53 |
| 1E+04 | 1E+05 | GABP01109393 | FSNMQTNLNRSIR | Phospho (ST)-S11 | P type ATPase | 18.25 |
| 1E+04 | 1E+06 | GABP01084447 | FTASRCFLMGSSSNAWA | Oxidation (M)- M10 Phospho (ST)-S12- S13-S14 | Histidine phosphatase superfamily clade 2 | 15.23 |
| 1E+04 | 1E+05 | GABP01044634 | FTSVKAVSMVK | Phospho (ST)-S3 | HSP | 22.6 |
| 1E+04 | 4E+05 | GABP01094306 | GAGSEGLLRPGPEAHATGQDPR | Phospho (ST)-S4 | 5-nucleotidase similar to putative chromatin structure regulator | 20.48 |
| 1E+04 | 2E+05 | JO757628 | GAPMTPTSMQGTVGSAPR | Phospho (ST)-T5 | | 79.01 |
| 1E+04 | 9E+04 | GABP01017523 | GEHPAHAPLLATGGYPPQAPSSGELQ | Deamidated [474]- Q18 Phospho (ST)-S22 | ubiquitin carboxyl-terminal hydrolase 20 | 16.75 |
| 1E+04 | 8E+04 | GABP01096263 | GESPVPVGAGEDQAHAK | Phospho (ST)-S3 | Fumarate hydratase | 40.35 |
| 1E+04 | 2E+05 | JO729762 | GGGCLSGDEGTPVASTRPLPSLGDPQ LR | Phospho (ST)-S6 | trehalose-6-phosphate synthase | 48.23 |
| 1E+04 | 3E+05 | GABP01015624 | GISTPHGIGTPGLGTGTPGVGTR | Phospho (ST)-T4 | Splicing factor 3B subunit | 21.14 |
| 1E+04 | 8E+05 | GABP01014465 | GMASAVADSFTSRDEAK | Phospho (ST)-S13 | step II splicing factor SLU7 | 18.83 |
| 1E+04 | 2E+05 | GABP01024097 | GPAPPPARAAGSGR | Phospho (ST)-S12 | Unknown | 28.89 |
| 1E+04 | 4E+06 | GABP01035382 | GPLFPGSSCFPLSPDHK | Phospho (ST)-S13 | MAPK | 67.32 |
| 1E+04 | 2E+06 | GABP01036082 | GPPPAPALRSR | Phospho (ST)-S10 | Unknown | 19.91 |
| 1E+04 | 8E+05 | GABP01060337 | GPSNRQHAGAGAAAPQPTR | Deamidated [474]- N4 Phospho (ST)-T17 | Unknown | 23.33 |
| 1E+04 | 7E+06 | JO710748 | GSFAPRTPGSPSPK | Phospho (ST)-S12 | starch synthase isoform I | 21.2 |

| | | | | | | |
|-------|-------|--------------|----------------------------|--|---|-------|
| 1E+04 | 3E+05 | GABP01047073 | GVVPEYSRPGSAYVVTR | Phospho (ST)-S11 | Ankyrin repeat protein | 24.84 |
| 1E+04 | 1E+05 | GABP01076526 | HADRCMAATSSQTNAIVGQSKPVAIL | Phospho (ST)-T9 -S10-S30 | BFR1 nuclear segregation protein | 18.42 |
| 1E+04 | 1E+05 | GABP01058092 | HLGLTPASPSSEQLGAPVFGSR | Phospho (ST)-S8 | Unknown | 39.16 |
| 1E+04 | 2E+05 | JO747619 | HPAGEVFLPNLPHDTSDDGYEMDG | Phospho (ST)-S17 | Unknown | 56.57 |
| 1E+04 | 1E+05 | JO734709 | PR | Phospho (ST)-S3 | Unknown | 44.78 |
| 1E+04 | 2E+06 | GABP01093073 | HRSPGAEDAPQPR | Phospho (ST)-S3 | Unknown | 42.63 |
| 1E+04 | 6E+05 | GABP01080080 | IPSPVALANLK | Phospho (ST)-S23 | ADP ribosylation factor like protein | 49.79 |
| 1E+04 | 4E+04 | GABP01047915 | IYHDEDEEGEERPDLDLLVVSSPK | Phospho (ST)-S15 | Unknown | 28.48 |
| 1E+04 | 2E+05 | JO723516 | KPAQEGAGAHPGHASAGEGGGATT | Phospho (ST)-S9 | ATP-dependent RNA helicase DBP3 | 40.8 |
| 1E+04 | 1E+05 | JO710317 | KVHTTFADSDDE | Phospho (ST)-S5 | erythrocyte-binding protein MAEBL-like | 15.67 |
| 1E+04 | 9E+05 | GABP01026003 | LASASVSSAWR | Phospho (ST)-S4 | CDC 37 like (protein kinase binding) | 44.62 |
| 1E+04 | 2E+06 | GABP01029881 | LELSDDEETFHPNLDK | Phospho (ST)-T6 | Aminoacylase 1 | 19.24 |
| 1E+04 | 7E+05 | GABP01042913 | LGDTVTVNLTMLR | Phospho (ST)-S7 Phospho (ST)-S11 | RNA pol II associated protein3 | 37.85 |
| 1E+04 | 2E+06 | GABP01019831 | LHIVEASDDDESDEEHNLPVR | Phospho (ST)-T8 | TPR domain protein | 16.45 |
| 1E+04 | 7E+04 | JO758130 | LLGSLFLTR | Phospho (ST)-T11 | Unknown | 27.78 |
| 1E+04 | 3E+05 | GABP01086664 | LLLLAAPPAATPSR | Phospho (ST)-S3 Phospho (ST)-S7 | PK domain protein | 17.69 |
| 1E+04 | 1E+05 | JO757521 | LLSPPPSPDTR | Phospho (ST)-S11 | TPR-repeat protein | 83.4 |
| 1E+04 | 4E+05 | JO734389 | LQQPSFLK | Deamidated [474]-Q2 Deamidated [474]-Q3 Phospho (ST)-S5 | ketoyl reductase domain protein | 22.78 |
| 1E+04 | 1E+05 | GABP01070161 | LSHRNSFCLQSSR | Phospho (ST)-S6 | EF hand-SH3 domain | 19.36 |
| 1E+04 | 5E+04 | JO719242 | MANSLSYMSAMPSSSQGCCPVR | Oxidation (M)-M1 Phospho (ST)-S4-S13-S14-S15 | similar to PKG | 16.6 |
| 1E+04 | 4E+05 | GABP01058002 | MEGDDIGHATSDSVLAR | Phospho (ST)-S11 | Vacuolar ion transport | 68.08 |
| 1E+04 | 3E+05 | JO744440 | NFSQSLICVDEAPAEAAAAAPSPK | Phospho (ST)-S3 Phospho (ST)-S24 | Soluble starch synthase 1, chloroplastic/amyloplastic | 50.49 |
| 1E+04 | 2E+05 | GABP01074076 | NMGSGMGSLSQSPAGAAK | Phospho (ST)-S11 | Unknown | 35.31 |
| 1E+04 | 4E+04 | GABP01023642 | NRGDDISEGGR | Phospho (ST)-S7 | Unknown | 29.98 |
| 1E+04 | 4E+05 | GABP01001808 | NSSYSCLSFLEGPNEGSEEPQHR | Phospho (ST)-S3 | starch synthase isoform I | 74.19 |
| 1E+04 | 3E+05 | GABP01113514 | PCGNMSVLPIGIMDMTFCDPSFVTY | Phospho (ST)-S28 Phospho (ST)-S30 | SRP54, signal recognition 54, GTPase. | 16.96 |
| 1E+04 | 1E+05 | GABP01043652 | CSCSYMSR | Carboxymethyl (C)-C8 Phospho (ST)-S13 | Ketol acid reductoisomerase | 16.07 |
| 1E+04 | 2E+06 | GABP01046424 | PCSQSPACSGPDSGPAR | Deamidated [474]-N3 Phospho (ST)-S11 Phospho (ST)-S15 | acyltransferase | 44.1 |
| 1E+04 | 4E+05 | GABP01050590 | PGNASPSPKGSQAGSGR | Deamidated [474]-Q1 Phospho (ST)-T13 Oxidation (M)-M17 | 50S ribosomal protein L5 | 17.92 |
| 1E+04 | 1E+05 | GABP01020469 | QPKAATLLLAFTTFGCMPEMR | Deamidated [474]-Q1 Phospho (ST)-T9 | Unknown | 34.18 |
| 1E+04 | 7E+04 | GABP01013760 | QKPLQLLVTR | Deamidated [474]-Q1 Phospho (ST)-S4 Deamidated [474]-Q14 | karyopherin beta, putative | 18.26 |
| 1E+04 | 3E+05 | GABP01038238 | QSPSPAPGPPAVVQAR | Phospho (ST)-S3 Phospho (ST)-S10 Deamidated [474]-Q13 | histidyl tRNA synthetase | 15.57 |
| 1E+04 | 2E+06 | GABP01080861 | RGSDPLPGDSPMQGIAALGGR | Phospho (ST)-S3 | Myotubularin related | 31.48 |
| 1E+04 | 1E+05 | GABP01061140 | RNSMEGTELK | Phospho (ST)-S5-S11-S12 | Unknown | 21.12 |
| 1E+04 | 1E+05 | GABP01061140 | RPRGSGPSAGSSR | | | |

| | | | | | | |
|-------|-------|--------------|--|--------------------------------------|---|--------|
| 1E+04 | 3E+05 | JO711998 | SCISPSTTSPCTTPMSMR | Phospho (ST)-S1-S4-S6 | Unknown 5-AMP-activated protein | 17.09 |
| 1E+04 | 3E+05 | JO706022 | SGQEFHVQELPR | Phospho (ST)-S1 | kinase , beta subunit | 40.93 |
| 1E+04 | 3E+05 | JO706757 | SKDADDSEEEEEEEGR | Phospho (ST)-S7 Phospho (ST)-S11 | Unknown | 77.45 |
| 1E+04 | 4E+05 | GABP01053461 | SLRSLWSLWLLPPR | Phospho (ST)-S1 | Unknown | 15.38 |
| 1E+04 | 5E+05 | GABP01107775 | SPAVVTTTGGSPVVMR | Phospho (ST)-S11 | Unknown | 68 |
| 1E+04 | 4E+05 | GABP01068339 | SRASAPCPAVLLCSRPGASSAALPR | Phospho (ST)-S14 | Unknown | 28.22 |
| 1E+04 | 7E+05 | JO726181 | SRTDSSGAAGNLER | Phospho (ST)-S6 Deamidated [474]-N11 | MAPK2, putative Sodium/potassium-transporting ATPase subunit | 23.18 |
| 1E+04 | 6E+05 | JO717713 | SSSATTRTWAMVATCSR | Phospho (ST)-S1-T14-S16 | alpha | 15.03 |
| 1E+04 | 4E+05 | JO695170 | SSSGLEDLQVEDGDGGSKK | Phospho (ST)-S1 | KH domain putative lactaldehyde | 67.5 |
| 1E+04 | 4E+06 | JO738704 | TASGAELAMEAEK | Phospho (ST)-S3 | reductase | 59.09 |
| 1E+04 | 4E+04 | GABP01043892 | TCFTTRIISAMFNTVTK TPGGWPSRPWPAWATGSPRSSPRP | Phospho (ST)-T14 Phospho (ST)-T16 | UDP-glucose dehydrogenase | 17.34 |
| 1E+04 | 5E+05 | GABP01029881 | R | Phospho (ST)-S19-S22-S23 | Aminoacylase 1 | 19.84 |
| 1E+04 | 2E+05 | GABP01096787 | TPSAALPMPAGEASGSGDGSPKK TQMHLHMDVILIFQRGNLSLDAGAA | Phospho (ST)-S20 | RRM-KH domain protein | 59.11 |
| 1E+04 | 7E+05 | GABP01093468 | DSSAVAGCVVSAR | Phospho (ST)-T1 | Unknown | 15.56 |
| 1E+04 | 5E+05 | GABP01026990 | TSQPTSLK | Phospho (ST)-S6 | Unknown | 20.53 |
| 1E+04 | 1E+06 | JO721047 | VLQLAQASR | Phospho (ST)-S8 | SET domain protein | 19.2 |
| 1E+04 | 8E+04 | JO753235 | VLVSQISEPAVKYSLWMSTMSSGLL K | Phospho (ST)-S4-S7-S14-S18-S22 | vitamin B12 dependent methionine synthase | 18.6 |
| 1E+04 | 9E+05 | GABP01098975 | VTASAMLPFERR | Phospho (ST)-S4 | Exonuclease/endonuclease/phosphatase eukaryotic translation | 15.14 |
| 1E+04 | 1E+05 | JO721439 | WADVDEEDEEGFNESPK | Phospho (ST)-S15 | initiation factor 3 p42 subunit | 112.11 |
| 1E+04 | 3E+05 | GABP01028262 | WPGRPGSAAGTRPGASWGR | Phospho (ST)-S16 | RRM domain protein | 16.26 |
| 1E+04 | 1E+06 | GABP01085452 | WSVRATPSVATAGR | Phospho (ST)-T11 | Na/H antiporter | 17 |
| 1E+04 | 9E+05 | GABP01073876 | WTSPGPSAR | Phospho (ST)-S3 | Phosphorylase kinase | 17.56 |
| 1E+04 | 1E+05 | GABP01005218 | YDEDTADEEVDER | Phospho (ST)-T5 | NTF2 superfamily | 45.28 |
| 1E+04 | 4E+05 | JO701316 | YGVVGQNGAGKTTLMK | Phospho (Y)-Y1 | translational activator | 16.07 |
| 1E+04 | 2E+06 | GABP01005692 | YLRQRPK | Phospho (Y)-Y1 Deamidated [474]-Q4 | Unknown | 34.25 |

4.6. Acknowledgements

We thank Eric Bonneil at the Université de Montréal's l'Institut de Recherche en Immunologie et en Cancérologie for protein sequencing and Mathieu Beauchemin for helpful discussions. Research support from the National Science and Engineering Research Council (NSERC) of Canada to DM (Grant number 171382-03) is gratefully acknowledged.

CHAPTER 5 – PUBLICATION # 4

Cold-induced cysts of the dinoflagellate *Lingulodinium* have low levels of protein phosphorylation and lack a normal circadian bioluminescence rhythm

Sougata Roy¹, Louis Letourneau² and David Morse^{1,3} (Submitted)

¹*Institut de Recherche en Biologie Végétale*

Département de Sciences Biologiques, Université de Montréal

4101 Sherbrookeest, Montréal, Québec, Canada H1X 2B2

² *Centre d'InnovationGénome Québec, McGill University,*

740 Docteur Penfield, Montréal, Québec H3A 1A4

I performed all the biochemical experiments and the fluorescence imaging in this section and drafted the first version of the paper. Louis Letourneau did the read assembly and the mapping of reads.

5.1. Abstract

Dinoflagellates are microscopic, eukaryotic and primarily marine planktons. Temporary cyst formation is a well-known physiological response of dinoflagellate cells to environmental stresses. However, the molecular underpinnings of cold-induced cyst physiology have never been described. Cultures of the dinoflagellate *Lingulodinium polyedrum* readily form temporary cysts when placed at low ($8 \pm 1^\circ\text{C}$) temperature and excyst to form normal motile cells following a return to normal temperature ($18 \pm 1^\circ\text{C}$). The circadian clock appears to be arrested in *Lingulodinium* cysts as the normal daily rhythm of Luciferin Binding Protein abundance is lost and after excystment the bioluminescence rhythm initiates at a time corresponding to ZT12 no matter when the cells were encysted. Phosphoprotein staining after 2D PAGE and column-based phosphoprotein enrichment followed by LC-MS/MS showed cyst proteins are hypophosphorylated when compared to those from motile cells, with the most marked decreases found for predicted Casein Kinase 2 (CK2) target sites. In contrast to the phosphoproteome, the cyst proteome is not markedly different from motile cells as assessed by 2D PAGE. RNA-Seq revealed cysts show a significant decrease in the levels of 132 RNAs. Of the 42 differentially expressed RNAs that were identified by sequence analysis, 21 correspond to plastid-encoded gene products and 11 to nuclear-encoded cell wall/plasma membrane components. Our data is consistent with a model in which the highly reduced metabolism in cysts is achieved primarily by alterations in the phosphoproteome. The stalling of the circadian clock suggests temporary cysts may provide an interesting model to address the circadian system of dinoflagellates.

Key Words: Cyst, Circadian clock, Post-Translational modification, Phosphoproteomics, Dinoflagellate, *Lingulodinium*

5.2. Introduction

Dinoflagellates are a group of unicellular and generally marine protists most closely related to the apicomplexans and the ciliates. They are known to contain large amounts of DNA [97] with a high proportion of unusual bases [121, 212], which is not organized into chromatin [184, 527] as there are no detectable histone proteins [188, 528], though mRNAs of all the components required to manufacture and modify nucleosomes were identified [102, 189]. They are major contributors to global primary production [131], promote biodiversity through their symbiosis with anthozoans in coral reefs [132], and can form the harmful algal blooms (HAB) commonly called “red tides” [529]. These HABs have been under tremendous scrutiny because of their ability to cause huge negative impacts on human health and marine based economy [530]. HAB formation is poorly understood, but while temperature and nutrient availability surely play an important role, the presence of cysts that can act as a reservoir for new populations may also be involved [531].

Dinoflagellate cysts are specialized cells with metabolism sufficiently reduced to enable them to resist poor environmental conditions; indeed, normal viable cells have been shown to emerge from cysts found in century-old sediments [532]. For *Lingulodinium polyedrum*, two types of cysts have been observed, and these are termed temporary (also asexual, ecdysal, or pellicle) and permanent (also sexual) cysts [533]. Permanent cysts are covered with numerous bulbous spines, and are thought to form part of the normal sexual life cycle. However, they have only been observed under laboratory conditions with some strains [534]. In contrast, temporary cysts form readily under a variety of conditions that constitute an environmental stress such as mechanical shock, changes in temperature, pH or salinity [130]. Interestingly, temporary cysts have also been observed to form after indolamine (melatonin) treatment or changes in photoperiod, providing an intriguing connection to the biological clock [535, 536]. The morphology of temporary cysts is distinct from that of the permanent cysts. Temporary cysts form by shedding their cellulosic thecal plates, which typically break along the groove in which the transverse flagella lies (the cingulum). The cysts that emerge are roughly spherical and are often seen as covered with a clear layer

[537]. The change in cell shape and in particular, the shedding off of the theca and the loss of flagella, is a common feature of temporary cyst formation and is indicative of intensive cytoskeletal rearrangements [538]. Temperature acts as a important cue for permanent as well as temporary cyst formation in dinoflagellates [537, 538]. Cyst formation strategy is employed by dinoflagellates to overwinter and bloom when temperature returns to normal [539] showing that in both types of cyst there is a substantial reduction in basic metabolism.

Coping with a cold temperature shock elicits similar responses in most living organisms, notably an extensive reorganization in the level of gene expression. Due to their greater complexity, higher organisms typically use cascades of changes in several regulatory pathways to achieve this, while single-celled organisms characteristically use more direct methods [540]. Extensive studies affecting gene expression due to cold temperature have been documented in plants and bacteria. Plants exposed to cold temperatures activate several factors including CBF (C-repeat/dehydration-responsive element binding factor) transcription factors, which in turn modulate the global stress-response transcription pattern [541]. Cold shock in bacteria induces a distinct set of proteins, comprised mainly of cold shock proteins (CSP) that, along with some helicases, nucleases, and ribosome-associated components, seem to be involved in nucleic acid metabolism. This response is specific to cold stress and does not correspond to other stress responses. Also, post-transcriptional events play an important role in bacteria [540]. Interestingly, *Lingulodinium* contains an unusually large number of cold shock domain proteins [165] although a role of these proteins in cold shock has not been previously examined.

Many physiological activities in *Lingulodinium* are under the control of an endogenous circadian (daily) clock [542]. This clock thus orchestrates circadian rhythms, the most studied of which is bioluminescence [79, 543, 544]. To produce changes in the bioluminescence capacity, the clock regulates translation of mRNAs encoding two key components required for light production (luciferase and a luciferin binding protein LBP), and levels of these proteins correlate with bioluminescent capacity, with maximum abundance at night and least at day [77, 78]. Daily synthesis of LBP at around dusk and its selective degradation during day has been used as a marker of the *Lingulodinium* circadian

clock. It is not known how cold-induced cyst formation in *Lingulodinium* affects the endogenous daily clock, although permanent cysts have been shown to contain a functioning yearly clock allowing seasonal excystment [545]. Several studies in diverse organism showed that their internal clock tends to hold around ZT12 when they were subjected to low temperature treatments [546]. In *Arabidopsis*, circadian oscillators demonstrated differential response under cold, and some cold-responsive genes were clearly controlled by the circadian clock [547]. Furthermore, cold temperature regulates alternative splicing of the CIRCADIAN CLOCK-ASSOCIATED1 (CCA1) mRNA, an important clock component in *Arabidopsis*, thereby modulating the clock function under cold stress [548].

To begin to understand the molecular basis for the profound cellular changes resulting in cyst formation, we have undertaken microscopic as well as proteomic and transcriptomic studies on the temporary cysts formed by *L. polyedrum* in response to low temperature. While the 2D PAGE protein patterns of cysts do not differ from motile cells, a major alteration was seen in phosphoprotein profiles. The low temperatures also appear to either stop the biological clock or decouple the clock control over translation as the bioluminescence rhythm in excysted cells is the same and is independent of the time of encystment and at the same time circadian synthesis and degradation of LBP is stopped in the encysted cells. Curiously, changes are also observed in the levels of some RNAs, especially those encoded by the plastid and those whose products are directed to the plasma membrane. We suggest these latter changes reflect selective RNA degradation and are a consequence rather than a cause of encystment.

5.3. Materials and Methods

5.3.1. Cell Culture

Lingulodinium polyedrum (formerly *Gonyaulax polyedra*; strain CCMP1936) was obtained from the Provasoli-Guillard National Center for Marine Algae (Boothbay Harbor, Maine). Cultures were grown in f/2 seawater medium lacking Si was used to culture the cells at $19 \pm 1^\circ\text{C}$ under 12-h light/12-h dark cycles using cool white fluorescent light at an intensity of $50 \mu\text{mol photons m}^{-2} \text{ s}^{-1}$. Under this light schedule, ZT0 (Zeitgeber Time) corresponds to the beginning of the light period and ZT12 the beginning of the dark period. Cultures were allowed to grow to a cell density of roughly 10^4 cells/mL and then harvested by filtration on a Whatman 541 paper supported by a Buchner funnel. The cells were either used immediately or stored at -80°C until further use.

5.3.2. Cyst formation and purification

We used temperature as a reliable and rapid method to elicit cyst formation by *L. polyedrum*. Cell cultures with an approximate density of 10^4 cells/mL were placed in a refrigerator maintained at $8 \pm 1^\circ\text{C}$. After 24 hours, the cysts and cell wall debris were collected by centrifugation (500 xg for 1 min in a swinging bucket rotor). After resuspension in seawater, the sample was layered on top of a 60% solution of Percoll (GE Healthcare) in seawater and again centrifuged (3000 rpm for 30 min in a swinging bucket rotor). Debris, mostly thecal plates as assessed by microscopic examination, settled at the bottom while healthy cysts remained at the interface. The cysts were collected and the Percoll removed by washing several times with either normal seawater (for samples used for SDS-PAGE) or phosphate-free seawater (for samples used for phosphoprotein purification). These cyst cells were used immediately for protein or phosphoprotein preparations. For some experiment cells from ZT6 and ZT18 were placed in the $8 \pm 1^\circ\text{C}$ refrigerator and samples were collected after 12 hours of incubation.

5.3.3. RNA Extraction and sequencing

L. polyedrum cysts were harvested after 24 hour incubation at $8 \pm 1^{\circ}\text{C}$. Cells were concentrated by centrifugation ($500 \times g$ for 1 min), washed with fresh seawater, and recentrifuged to reduce bacterial contamination. Total RNA was isolated with TRIzol (Invitrogen), as per the manufacturers' protocol. RNA samples were subjected to quality control assessment using a Bioanalyzer (Agilent), and sequencing used an mRNA-Seq sample preparation kit (RS-100-0801 from Illumina). Sample preparation for RNA Seq and HiSeq Illumina sequencing was at the McGill University and Génome Québec Innovation Centre (Montreal, Quebec). Roughly 89 million 100 bp paired end reads (~18 Gbp) were obtained from the cyst sample and assembled together with 312 million 76 bp PE reads from non-encysted cells [165] using Trinity [549]. The 2 first steps of Trinity [550], inchworm and chrysalis, ran with default parameters. The last step, butterfly, was split to run in parallel on a cluster. The final assembly containing 114,779 sequences of ≥ 300 bp has been deposited in GenBank (Transcriptome Shotgun Assembly project accession number GABP01000000).

5.3.4. Sequence Analysis

Cyst reads were mapped to the previous Velvet assembly [165] and to the new Trinity assembly using BWA [551], and compared to similar mappings using RNA from motile cells harvested at ZT 18 and CT 18 (accession numbers SRR330444 (ZT18), SRR330445 (CT18) and SRR584359 (Cyst)). Read counts were analyzed by DegSeq to uncover statistically significant differences [552].

5.3.5. Northern hybridization

To verify the decrease of plastid transcripts in cyst cells, RNA was extracted from ZT6 and a 4 hour cold-incubated *Lingulodinium* cell sample. RNA was dissolved in DEPC treated water and the concentrations were estimated using a Nanophotometer (MBI). 10ug (1X) of ZT6- or 1ug (0.1X), 10 ug (1X) and 20 ug (2X) cyst RNA was adjusted in RNA

sample buffer (2.2M Formaldehyde, 50%formamide, 0.5X MOPS), heated at 55 °C for 15 minutes, chilled in ice for 1 minute and then mixed with 3uL of 10X RNA loading dye. 33ul of each sample was then loaded in a MOPS-formaldehyde-1% agarose gel and after completion of electrophoresis was washed for 15 minutes in 10X SSC buffer. RNA was transferred for overnight on to Nytran supercharge nylon membrane (Schleicher &Schuell BioScience GmbH) using 10X SSC buffer. After the blotting reaction, the RNA was crosslinked using a UV-crosslinker (Hoefer), prewashed in tubes for 2hours at 65 °C in RNA hybridization buffer (0.5M phosphate buffer pH 7.2, 1mM EDTA, 7% (w/v) SDS and 1% (w/v) bovine serum albumin) in a HB-1000 Hybridizer (hybridization oven from UVP) with rotation at 65 °C. Probes were prepared from *Lingulodinium* atpB and psbC (plastid RNAs), PCP (a nuclear-encoded plastid directed protein) and rRNA (control) using the Prime-a-gene labeling system (Promega) based on the random priming principle [553] and then purified by size exclusion chromatography on a biogel P60 column. The hybridization with radiolabelled probes was carried out for overnight at 65°C, after the hybridization reaction the membrane was washed sequentially with washing buffer 1 (40mM phosphate, 5% SDS, 1mM EDTA) and 2 (40mM phosphate, 1% SDS, 1mM EDTA) respectively at 65 °C for 15 minutes each. The blot was then exposed to the PhosphorImager screen for overnight. A typhoon 9200 PhosphorImager (GE Healthcare) was used to capture the radioactive emission. The same membrane was stripped using the hot SDS procedure (Amersham) and used for four sequential northern hybridization reactions.

5.3.6. Microscopy

Samples of both cysts and normal day phase *L. polyedrum* cells were imaged using an Axio Imager microscope equipped with epifluorescence (Zeiss). Scintillons (the bioluminescent organelles) were visualized using a CFP filter set (Ex: 485/20 nm; Em: >470 nm) while a Propidium Iodide filter set (Ex: 560/40 nm; Em: 630/75 nm) was used for detection of chlorophyll fluorescence. The sequential pictures of normal bright field image were taken along with the fluorescence images. An FEI Quanta 200 3D (Dualbeam) was used for the environmental scanning electron microscopy (SEM) of the *L. polyedrum* cysts. For the SEM and the fluorescence images, either 1 mL sample of cyst cells from a 24-hour

cold-incubated culture or normal cells was concentrated to 200 μ L and a 20 μ L aliquot was placed directly under the microscope without any further treatment.

5.3.7. Protein and phosphoprotein extraction

For 2-D PAGE experiments, both 24-hr cyst samples as well as samples taken from early day (ZT2) and early night (ZT14) normally growing *L. polyedrum* cultures were used for protein extractions. For 2-D gel electrophoresis analysis, total proteins were isolated using TRIzol (Invitrogen), using manufacturers' protocol with slight modifications. Briefly, 1 mL of TRIzol was added to 0.1- 0.2 g wet weight of cells, and after adding zirconium beads, cells were broken in a BeadBeater (BioSpec Products) at 4°C with two 1 min treatments. The lysate was incubated for 5 minutes at room temperature (RT) and then centrifuged at 12,000 x g for 10 min to remove debris. To the supernatant was added 200 μ L of chloroform, and after vortexing for 15 seconds the sample was left for 5 minutes at RT. The top aqueous phase containing the RNA was discarded and 300 μ L of 100% ethanol was added to the lower organic phase, mixed well and incubated for 5 minutes at RT. The samples were then centrifuged for 5 minutes at 2000 x g at 4°C to remove the DNA. The supernatant was distributed equally in two 2 mL Eppendorf tubes and 750 μ L of isopropanol was added to each, after which the samples were mixed well and then incubated at RT for 30 minutes. The total protein was precipitated by centrifugation at 12,000 x g for 10 minutes at 4°C. The precipitated proteins were washed twice with 1 mL of 95% ethanol for 20 minutes each and centrifuged at 7,500 x g for 5 minutes at 4°C after each wash. The samples were allowed to dry for 10 min at RT.

For SDS-PAGE and immunoblotting, filtered cells and cysts were resuspended in a protein extraction buffer (10 mM Tris-HCl pH 8.0; 20 mM NaCl; 1 mM DTT; 1 mM EDTA; 0.1% NP40; 1 mM PMSF supplemented with 1X protease inhibitor) with zirconium beads. The cells were broken with two 1 minute treatments in the BeadBeater and insoluble cell debris removed by two sequential centrifugations at 11,000 x g for 10 and 5 minutes respectively, each at 4°C. Protein concentrations in the supernatant were measured using Bradford assay (BioRad) in a VersaMax (Molecular Devices) plate reader.

A commercial phosphoprotein purification kit (Qiagen) was used to enrich for phosphoproteins. Washed cysts and cells were crushed in liquid nitrogen using a mortar and pestle, immediately added to the lysis buffer supplied by the manufacturer, and incubated on ice for 30 min with occasional mixing. Insoluble cell debris was removed by two sequential centrifugations at 15,000 x g for 10 and 5 minutes respectively, each at 4°C, and the supernatant retained. Protein concentrations were measured using Bradford assay as above. Total protein (2.5 mg) was diluted to a final concentration of 0.1 mg/mL in 25 mL of lysis buffer supplemented with benzonase and protease inhibitors and passed through the affinity column at room temperature to capture the phosphoproteins. Unbound proteins were removed by washing with the lysis/wash buffer after which phosphoproteins were eluted with the supplied elution buffer. The eluted phosphoproteins were desalted and concentrated by centrifugation on a amicon ultra-4 (Millipore) and precipitated with 4 volumes of prechilled (-20°C) acetone for 2 hours at -20°C.

5.3.8. 2-D gel electrophoresis

For isoelectric focusing (IEF), a dry protein pellet (from trizol method) was resuspended in 200 µL lysis buffer containing 7 M urea, 2 M thiourea, 4% CHAPS, 0.001% Bromophenol blue supplemented with 20 mM dithiothreitol, and 0.1% BioLyte pH 3-10 ampholyte buffer (GE Healthcare). This total protein sample (0.5 mg in 125 µL of lysis buffer) was used to rehydrate a 7 cm pH 4-7 nonlinear immobilized pH gradient (IPG) strips (Bio-Rad). After 12 hours of rehydration the IEF was carried out as per manufacturer's protocol. Strips were then prepared for the second dimension by a 10 min incubation in 0.375 M Tris-HCl (pH 8.8) containing 8 M urea, 2% SDS, 20% glycerol, 2% DTT followed by a 10 min incubation in the same buffer containing 2.5% iodoacetamide. The strip was then soaked briefly in 1X Tris-Glycine buffer, placed on top of a 12.5% polyacrylamide gel containing SDS, and overlaid with 0.25% agarose in Tris-Glycine buffer containing trace amounts of Bromophenol blue. The second dimension was electrophoresed as for single dimension SDS gels. Gels were stained sequentially with ProQ Diamond (Invitrogen) and Coomassie Brilliant Blue R-250 [549]. A Typhoon 9200 PhosphorImager (GE Healthcare)

was used to visualize protein fluorescence while the Coomassie stained gels were scanned using an AGFA DuoScan T1200.

5.3.9. SDS-PAGE and Western blotting

Lingulodinium ZT2, ZT14 and cyst protein samples were dissolved in SDS sample buffer (2% SDS, 0.7 mM 2-mercaptoethanol, 62.5 mM Tris-HCl, pH 6.8, 10% glycerol) and heated at 95 °C for 5 min. 30 µg of proteins from each sample were then resolved by SDS-PAGE on 12.5% polyacrylamide gels. Precision Plus Protein Standards (Bio-Rad) were used as molecular weight markers. For Immunoblotting, the proteins from gels were transferred to the Hybond-P PVDF membranes (Amersham Biosciences) using the Transblot SD Semi-Dry Electrophoretic transfer cell (Bio-Rad) following the manufacturer's protocol. After blocking the membranes with 5% Non-fat dry milk in Tris-buffered saline buffer supplemented with 0.05% Tween-20, immuno-reaction was performed with custom-made rabbit polyclonal antibodies raised against Peridinin-Chlorophyll *a*-binding Protein (PCP; 1:100,000), Luciferin Binding Protein (LBP; 1:5,000) and Ribulose biphosphate Carboxylase/Oxygenase (Rubisco; 1:150,000) in the same buffer. After reaction with a secondary antibody (1:30,000) and subsequent washings, the blots were developed with chemiluminescent substrate (Millipore) and were exposed to the ImageQuant LAS 4000 (GE Healthcare) to capture the chemiluminescence.

5.3.10. Mass spectrometry analysis

The acetone precipitated enriched phosphoprotein samples from *Lingulodinium* were used directly for trypsin digestion and long run LC-MS/MS analysis at the proteomic facility of l'Institut de Recherche en Immunologie et en Cancérologie (IRIC, Université de Montréal, Canada). Briefly, 20 µL of a trypsin digestin 5% acetonitrile/0.2% formic acid was injected on a C18 precolumn (0.3 mm i.d. x 5 mm) and peptides were separated on a C18 analytical column (150 µm i.d. x 100 mm) using an Eksigent nanoLC-2D system with a 76-min gradient from (A/B) 10–60% (A: formic acid 0.2 %, B: acetonitrile/0.2% formic acid). The LC system was coupled to a LTQ-OrbitrapVelos mass spectrometer (Thermo Fisher). Each

full MS spectrum was followed by 12 MS/MS spectra (thirteen scan events), and the 12 most abundant multiply charged ions selected for MS/MS sequencing. MS-MS was performed using collision-induced dissociation in the linear ion trap. Data were processed with the Mascot 2.2 (Matrix Science) search engine using both a previously described *L. polyedrum* transcriptome assembly [165] and the Trinity assembly. The variable modifications included were deamidation [474], carbamidomethylation (C), oxidation (M) and phosphorylation (STY). Precursor and fragment tolerances were 10 ppm and 0.5 Da, respectively.

Peptide abundances were compared between samples using raw data files (.raw) from the Xcalibur software, which were first converted into peptide map files representing all ions according to their corresponding m/z values, retention times, intensities, and charge states. Intensity values above a threshold of 10,000 counts were considered for further analysis. Peptide abundances were assessed using the peak top intensity values. Clustering of peptide maps across different sample sets were performed on peptides associated to a Mascot entry using hierarchical clustering with tolerances of 15 ppm and 1 min for peptide mass and retention time, respectively. Retention time of the initial peptide cluster list was normalized using a dynamic and nonlinear correction to confine the retention time distribution to less than 0.1 min (<0.3% RSD) on average.

The variation of intensities between samples was used to compute the fold change of a protein. First, a number between 0 and 1 that described the amount of representation of the protein within each condition was assigned. Then an in-house software (ProteoProfile; <http://www.thibault.irc.ca/proteoprofile/files/TechnicalGuide.pdf>), which assigns weights to the peptides composing the protein, was used to calculate the relative intensities for each protein. The weight of each peptide represents its potential to describe correctly the protein. Each peptide starts with a weight in proportion to its own intensity level (Log 10 of the average intensity of the peptide divided by 10). Based on the Weiszfeld's iteratively re-weighted least squares algorithm, this weight is multiplied by the closeness of the peptide to the protein's fold change, through a series of iterations.

5.3.11. Bioinformatic analysis

The LC-MS/MS analysis returned a list of peptides along with their relative intensities from ZT2, ZT14 and cyst samples. Some peptides appear several times in the list, therefore intensities of the peptides with identical sequences were summed to yield final peptide intensity. For some analyses, all peptides identified by comparison to the *Lingulodinium* transcriptome were used as all are potentially derived from phosphorylated proteins. However, for most analyses only peptides with an identified phosphosite(s) were used. We classified the cyst phosphopeptides as hyperphosphorylated when their intensity were 2-fold above the highest intensity observed in ZT2/ZT14 and hypophosphorylated when their intensity were 2-fold below the lowest observed intensity between ZT2 and 14.

Sequence annotation and mapping to Gene Ontology (GO) [441] and Interpro domains [475] were performed using the web based tool Blast2GO [443]. For annotating the Cyst hypo and hyper phosphorylated proteins, the Interpro domain information was used to infer a function for some sequences where a GO category for the protein itself was not available. Information from these two different sources was merged and verified manually to obtain the largest number possible of identified proteins.

5.3.12. Bioluminescence assay

To check the effect of cold treatment on the bioluminescence rhythm, two out of phase cultures (ZT4 and ZT16) were taken from two different culture rooms running simultaneously with opposing light dark regimes and placed in an $8 \pm 1^\circ\text{C}$ refrigerator in dim light. After 8 hours, the encysted cells were taken out and quadruplicate samples added to a 96 well microtiter plate along with cells from both culture rooms (now at a time corresponding to ZT12 and ZT24). The plate was placed in a microplate reader (Spectramax M5 from Molecular devices) kept at culture room conditions and the bioluminescence recorded each 2 minutes for the next 70 hours in constant dark. All samples were surrounded by wells containing only seawater to eliminate interference by bioluminescence from adjacent wells.

5.4. Results

5.4.1. Cold temperatures induce temporary cysts in *Lingulodinium*

Lingulodinium responds very rapidly to low temperature, as within 2-3 hours of incubation at $8 \pm 1^\circ\text{C}$, the cells no longer swim and have settled at the bottom of the flask. These cells shed their theca and flagella within the next few hours and have assumed the rounded shape surrounded by a thin pellicular layer consistent with the formation of temporary cysts (Fig. 5.1.A, B). After returning the cysts to the normal culture room temperature ($18 \pm 1^\circ\text{C}$), cells were found to have completely excysted and resumed swimming within few hours, indicating that the changes are fully reversible and relatively rapid. Interestingly, even after 24 hours at 8°C the cysts have retained scintillons, the bioluminescent organelles, in numbers similar to what are typically found in day phase *Lingulodinium* cells (Fig. 5.1.E, H). Chloroplasts are also retained in cysts, although they appear disorganized in the cytoplasm and display a more rounded morphology (Fig. 5.1.D, G). However, cells stained with DAPI then allowed to encyst showed no change in nuclear morphology (not shown).

5.4.2. Protein phosphorylation is reduced in cysts.

To address the molecular changes accompanying cyst formation in response to a cold shock, we first examined the protein profiles using 2D-PAGE (Fig. 5.2.). The general protein pattern revealed by Coomassie blue staining of total cyst proteins is indistinguishable from that obtained using motile cell proteins, taken either during the early day or the early night phase of the daily light-dark cycle. To assess the possibility that some proteins were transiently overexpressed during the process of encystment, 2D-PAGE analysis was also performed using proteins isolated from cells after 5 hours incubation at 8°C . However, no difference was observed in these protein profiles compared to the 24-hour cysts (not shown). To further confirm the 2D gel patterns, western blot analyses were performed using antibodies recognizing the proteins LBP, Rubisco and PCP, and these also showed no significant variation between cysts and motile cells (Supplementary Figure 5.S1.).

In contrast to the similarity in general proteins, the 2D gels of cyst and motile cell extracts show a marked difference after staining with the phosphoprotein stain ProQ diamond (Fig 5.2.). We note that while some individual proteins may have a greater staining intensity, the general overall staining intensity is lower when cyst extracts are observed. To provide a global characterization of the differences between cysts and motile cells, we prepared a phosphoprotein-enriched fraction from cysts as well as from cells harvested during early day (ZT2) and early night (ZT14). These phosphoprotein-enriched fractions were then digested with trypsin and analyzed by long run LC-MS/MS. Over 12,000 peptides were recovered from combined samples, of which 618 contained at least one phosphorylated amino acid as determined by mass spectroscopy. Only 23 of these phosphopeptides have a phosphotyrosine.

A comparison of the signal intensity of all 618 phosphopeptides between cysts and motile cells showed that three general classes could be discerned. A first class, containing almost half the phosphopeptides (306), was less abundant in cyst extracts than in either ZT2 or ZT14 extracts. In principal, the lower levels of these peptides could be due to a change in the phosphorylation state of the protein or a change in the amount of the phosphorylated protein. However, since the protein patterns of cysts and motile cells are similar, we term these peptides hypophosphorylated. Unfortunately, it is not possible to accurately measure the degree of hypophosphorylation for these peptides, as signal intensities below a threshold of 10,000 are all arbitrarily assigned the threshold value. An additional 173 peptides were above threshold levels in the cyst extracts but at threshold levels in either one or both motile cell extracts and are termed hyperphosphorylated. The remaining 141 phosphopeptides were above threshold in all three extracts and have a similar degree of phosphorylation in all. The over-representation of hypophosphorylated peptides is more clearly indicated by a plot of the ratio of signal intensities in cyst compared to motile cells for each of the individual peptides (Fig. 5.3.A). This figure displays the 123 peptides whose signal intensity is at least 2-fold more than the highest value in motile cells (i.e., in either ZT2 or ZT14) and the 242 peptides whose signal intensity is at least two-fold less than the lowest value in motile cells. The

prevalence of hypophosphorylation is clearly illustrated by the larger area under the lower curve.

The 618 phosphopeptides could be assigned to a total of 570 phosphoproteins, and of these, 117 and 221 proteins are hyper or hypophosphorylated in the cysts, respectively. These 338 phosphoproteins were annotated manually by combining identification from BLAST searches and protein domain information (Supplementary Table 5.ST1. and 5.ST2.), then classified into probable functional categories (Fig 5.3.B). The inclusion of information from protein domains increased the number of identified probable functions to roughly half of the proteins. The majority of these proteins were classified into categories encompassing general metabolism or the regulation of gene expression at the transcriptional, translational or post-translational levels, as expected based on the lower metabolism expected for cysts. The post-translational category contains proteins involved in ubiquitinylation, protein phosphorylation (kinases and phosphatases) and protein-protein interactions. The translation category contains proteins with RNA-binding domains, translation factors, tRNA biosynthesis and splicing factors. The transcription category contains proteins involved in transcription as well as other DNA binding protein, while the cytoplasm category contains enzymes catalyzing a wide variety of reactions for primary and secondary metabolism.

We also performed the cell component analysis of the global phosphoprotein enriched fraction from cysts and then compared them to that of the ZT2/14 motile cells. The maximum differences are found in the cytosol, mitochondrion, nucleus and ribosome compartments, indicating a similar post-translational control of gene regulation and metabolism pathway proteins by changes in phosphorylation (Supplementary figure 5.S3.).

5.4.3. Cysts have an arrested clock and show a decreased level of Casein Kinase 2 phosphosites

The amino acid context surrounding the phosphorylated amino acid allows the peptides to be categorized as potential targets for different kinases [480]. Using the

information of available kinases in *Lingulodinium* (Chapter 4), we used the GPS algorithm to predict kinase sites within the phosphopeptides obtained from MS study. Interestingly, this analysis indicates that predicted Casein Kinase 2 (CK2) phosphosites appear to be most affected by the general decrease in cyst protein phosphorylation (Fig 5.4.A). Indeed, the proportion of hypophosphorylated CK2 sites exceeds the average number of hypophosphorylated sites by roughly 50% (Fig 5.4.B). CK2 is a conserved kinase in the mechanism of eukaryotic circadian clocks [82, 91], so we therefore tested the impact of encystment on the clock. Cells from cultures grown under two light regimes 12 hours out of phase were encysted at the same time by low temperature (for eight hours) then allowed to excyst in a plate reader maintained at normal culture room temperature. Bioluminescence measurements show that rhythm in the two cultures is out of phase before encystment yet in phase after excystment (Figure 5.4.C). This result indicates that at excystment the circadian clock starts for both cultures at the same time (corresponding to the onset of the subjective dark phase). It is known that LBP synthesis is under the control of the circadian clock, so an easy way to test the status of the *Lingulodinium* bioluminescence clock under cold stress is to monitor the rhythmic levels of this protein. We performed this by taking a culture from midday (a time when LBP protein levels are low) or midnight (when LBP amount are the highest) and incubating it in cold (8°C). We collected the two cold-incubated samples after 12 hours and then compared their LBP status. Ideally, if the clock is still functional at low temperature, the LBP abundance of the two 12 hour cold-incubated samples should be opposite to the corresponding ZT times. However, Immunoblotting with LBP antibody showed similar protein levels between the motile cells and the out of phase 12 hour cold incubated cysts (Figure 5.4.D), indicating that the circadian clock is frozen in the cysts and hence cannot regulate the daily synthesis and degradation of LBP anymore.

5.4.4. Plastid-encoded RNAs have decreased levels in cysts

Transcription is known to be modulated under cold stress in plants, as is the case in *Arabidopsis* where 4–20% of the genome is demonstrated to be cold-regulated [554]. Therefore, as a complement to the examination of the proteome and the phosphoproteome, RNA-Seq was used to evaluate potential changes in cyst RNA levels. Read counts for each

of the 74,114 consensus sequences in a previously described assembly [165] were determined for RNA samples prepared either from cysts or from motile cells taken during the middle of the night phase (ZT18). Statistically significant differences were assessed using DegSeq, as illustrated using a comparison of read counts determined using cyst RNA and RNA from motile cells harvested during the middle of the night (Fig 5.5.A). This analysis plots the fold difference against mean transcript abundance for all sequences, and sequences with significantly different read counts are highlighted by colored points. Interestingly, the majority of the differences reflect a decrease in the amount of RNA in cyst compared to motile cells. The analysis was repeated with a replicate of ZT 18 sample from motile cells, and those sequences with significantly different values in both samples were retained for further analysis. Using this criterion, a total of 132 RNA were found to have a differential abundance in the cyst, 9 with higher values in the cyst and 123 with lower values in the cysts. Roughly 70% could not be identified, either because BLAST searches (BLASTx with an E-value $<e^{-05}$) did not return a match or because matches were only to unidentified proteins, a proportion similar to that found for the total transcriptome [165]. Interestingly, among the 42 sequences that could be identified, 19 highly expressed transcripts all encoded by the plastid genome (orange circles) were found to have decreased. We also verified this decrease of all plastid encoded RNAs using the same analysis pipeline but with a transcriptome assembled by Trinity. This decrease in the level of plastid-encoded RNAs was confirmed by Northern blots for *atpB* and *psbC* gene products (Supplementary Fig 5.S2.) in order to eliminate the possibility that this might be due to a technical artifact resulting from slight differences in the degree of purification of polyadenylated RNAs from cyst and motile cell RNA. Intriguingly, 11 transcripts representing nuclear-encoded plasma membrane proteins were also found to be present at lower levels in cysts. It is unclear how the decreased levels of transcripts encoding membrane proteins might affect the cellular structure, given that protein levels on 2D PAGE did not change markedly, although as yet we are unaware of the levels of these proteins. But it is clear that the appearance of the cell wall is markedly different in cysts. It seems likely that these differences in RNA levels can result from either an increase in specific degradation rates, a decrease in transcription rates or a combination of both.

5.5. Discussion

Encystment is a physiological response of many dinoflagellates to adverse environmental conditions associated with either abiotic agents as temperature [126], nutrients [555], indolamines [536] or biotic agents such as bacterial attack [129], culture age [128], competing phytoplankton [556] or predators [557]. Cold –induced cyst formation has also been reported in the dinoflagellate *Pfiesteria piscicida*, where the maximum excystment occurred between 9-18 hour after returning the encysted cells to the normal temperature and light regime [558]. Cyst formation requires an extensive structural reorganization, starting with the formation of a new cell wall replacing the original thecal plates. This new structure is thought to contain a dinoflagellate-specific version of sporopollenin, a highly resistant and presumably impermeable covering [559], protecting them from adverse environment as well as toxic chemicals that could be fatal to the normal cells. The impermeable nature of this covering will of necessity decrease gas exchange and thus general metabolism in the cysts. Interestingly, some studies reported the breakdown and alteration in the distribution pattern of pigments in temporary cysts [560], which might be correlated with the changes we observe in chloroplast organization (Fig 5.1.). However, unlike chloroplasts, the nuclear structure retains its typical shape. The cysts formed here by exposure to cold appear identical in morphology to those obtained from indolamine treatment [127] suggestive of a common mechanism initiated by a variety of stimuli. Surprisingly, even though temporary cyst formation under different biotic and abiotic environmental stress is common among dinoflagellates, the molecular basis of this response is not yet known.

Some aspects of cold shock response appear similar in several unrelated organisms such as bacteria, plants and higher eukaryotes. In particular, the rapid translation of CSPs (cold-shock domain containing proteins), a specialized group of multifunctional RNA binding proteins [561] appear important. The function of these CSPs during cold shock is to bind non-specifically to mRNA and relieve cold-induced secondary structure conformations [562] that could then modulate the levels of general protein synthesis. In *E. coli*, a prokaryotic model of the cold shock response, a shift in temperature from 37 to 10 °C shows

that cells specifically overexpress CSPs over a period of approximately 4 h, while the global transcription and translation rates decrease [563]. This is in sharp contrast to what has been observed in plants, where cold stress is accompanied by alterations in protein synthesis principally targeting photosynthesis and carbohydrate metabolism pathways [564-567]. Intriguingly, CSPs are the most abundant nucleic acid binding domain found in *Lingulodinium* [165] and *Symbiodinium* [102], yet no evidence is seen for an increase in protein abundance of any type on 2D-PAGE (Fig. 5.2.). Therefore, if dinoflagellate CSPs functions as part of the response to cold shock, levels of the proteins must be below detectable limits of the 2D gels. It is possible, however, that CSPs do not function in cold adaptation in dinoflagellates but act as regulators of gene expression during normal growth. This is supported by the observation that CSPs in plants play a role in many cellular processes under standard growth conditions [519] as well as by the fact that *Lingulodinium* in the cold prefers to slow down global metabolism by encysting rather than trying to maintain normal metabolism.

Cold acclimation in plants involves excessive reorganization of their transcriptome in response to cold stress [568, 569]. In contrast, the rapidity of temporary cyst formation in *Lingulodinium* and the ease with which this process can be reversed, argues against major changes in the transcriptional response of the cells. In addition, the fact that the majority of the changes seen in cyst transcript levels in response to cold treatment are reductions rather than increases in the cyst transcriptome also supports this idea. The decrease of specific RNA levels seen in *Lingulodinium* is similar to the decreased level of transcripts belonging to categories related to stress and hormonal response proteins in *Arabidopsis* [570], and suggests that in both cases a response to stress can involve targeted decay of specific mRNAs. However, *Lingulodinium* differs in that the two largest classes of RNAs identified included representatives from plastid-encoded and nuclear-encoded genes (Fig 5.5.B). It is likely that two mechanisms will have to be invoked to explain these different classes of RNA. The selective degradation of cytoplasmic RNAs has been well documented in prokaryotes and eukaryotes [570-572], and based on these systems, it is possible be that transcripts in *Lingulodinium* may be targeted for degradation by the presence of a common sequence motif or structural element such as the AUUUA motifs (AU-rich element) in the

3'UTR found in other systems [573-576]. Once the complete sequences of all the transcripts regulated in *Lingulodinium* cysts is known it will be possible to assess this possibility. In *Arabidopsis*, cold shock resulted in the decrease of mRNAs encoding pentatricopeptide (PPR) repeat family proteins [570]. PPR repeat proteins are nuclear-encoded proteins that are directed to organelles, where they are responsible for stabilizing organelle-encoded transcripts in plants and *Chlamydomonas* [577]. As yet, there is no such indication in *Lingulodinium*. Also, in *Lingulodinium* no nuclear-encoded mRNAs for plastid-directed proteins were specifically diminished under cold stress.

In addition to the changes in transcript abundance, substantial differences were also observed between the phosphoprotein profiles of cysts and motile cells. This aspect is similar to what has been observed for the phosphoprotein profile of rice plants and roots when cold stressed and normal plants were compared [578]. Phosphoprotein profiles of proteins involved in glycolytic pathway, carbohydrate metabolism, calcium mediated signal transduction and redox homeostasis differed under cold stress in the rice [578]. In alfalfa, prominent changes in phosphorylation pattern of the nuclear proteins were observed under cold shock (4°C) as compared to cells maintained at normal temperature [579]. These results are thus akin to what we observe for the phosphopeptide profiles in cysts, where levels of phosphoproteins involved in general as well as nucleic acid metabolism were found to differ markedly between cysts and motile cells. The general hypophosphorylation pattern observed in cysts (Figure 5.2. and Figure 5.4.B) might be due to the differential activity of kinases/phosphatases under cold shock, which might itself be regulated through phosphorylation events, as approximately 15% of the identified phosphoproteins in cysts are either kinases or phosphatases (Figure 5.3.). It seems *Lingulodinium* prefers to affect proteins post-translationally to modulate cellular activities under cold stress.

There are roughly 90 proteins in *Lingulodinium* cysts whose phosphorylation levels increase compared to motile cells. Among these, a putative phenylalanine tRNA synthetase is ~1000 times more phosphorylated in the cold-stressed cells. In mammals, phosphorylation of tRNA synthetases did not affect the amino-acylation activity but instead it enhanced the ability to synthesize di-adenosine tetraphosphate (Ap4A) by 2-6 fold [580], and yeast

phenylalanine tRNA synthetase has shown to synthesize Ap4A [581]. Ap4A has been called an alarmone [582], as its concentration increases during different stresses [474, 583], and it has been implicated in a variety of cellular activities including cell division, DNA polymerase activity, and the activation or inhibition of particular cellular enzymes [584]. It is an intriguing possibility that Ap4A may play a role in down-regulating the metabolic activities in *Lingulodinium* cysts. It would thus be of interest to test for the presence of Ap4A in cysts.

It is important to stress that measurements of phosphopeptide intensity alone cannot distinguish between changes in phosphorylation state of a constant amount of protein or changes in the absolute amounts of a protein whose phosphorylation state remains constant. However, the observation that the amount of those cyst proteins resolved by 2D PAGE does not differ from what is found in motile cells (Fig 5.2.) suggests that changes in phosphorylation state of different proteins may be the more likely scenario. This would also explain the rapidity of encystment and excystment as well as the observation that a sizeable fraction of the phosphopeptides whose levels change markedly in cysts can be classified into a kinase/phosphatase category (Fig 5.3.). To more provide more support for this idea, however, measurements of the amount of selected proteins should be tested by Western blots to allow precise comparison of the amount of the protein and the amount of the differentially phosphorylated form as 2D gels only represents the most abundant and stable proteins.

It is interesting that almost 8% of the identified proteins demonstrate some form of calcium or cAMP regulation (calcium binding and calcium or cAMP regulated kinases). Calcium and cAMP are well known second messengers and may be implicated in indolamine-induced cyst formation in dinoflagellates [127]. Studies in plants have demonstrated that Ca^{+2} signaling is an important plant response to several abiotic stresses including low temperature [585]. It is also of interest that putative CK2 kinase target motifs are enriched among the hypophosphorylated peptides. CK2 is conserved in circadian clock mechanisms of all eukaryotes [89, 472, 586] and it may thus be relevant that circadian clock function is demonstrably affected by cyst formation.

Interestingly, the cold treatment has a major impact on the circadian behavior of the cells. We find the bioluminescence rhythm of two originally out-of-phase cultures begins at the same phase after excystment (Fig 5.4.C), indicating that the bioluminescence clock starts from the same time (roughly ZT12) when cells are returned to normal culture room temperatures. This result could reflect a strong phase resetting of a clock that continues to function in the cyst or alternatively, the “rebooting” of a clock that had stopped in cysts. However, in support of the stopped clock, the amount of LBP, whose levels normally oscillate over the circadian cycle, is not rhythmic in cysts (Figure 5.4.D). Thus cells encysted during midday still have low LBP levels 12 hours later, when in motile cells the levels have increased substantially. Similarly, LBP levels in cells encysted at midnight remain high 12 hours later, even though levels have decreased in motile cells. The interpretation favored by these results is that the clock has stopped in cysts and restarts from a time corresponding to the levels of any clock components found normally at ZT12. This arrest of the circadian clock on temporary cysts thus appears different from the endogenous annual clock in permanent cysts which remains active in regulating cyst germination [557].

On a practical level, dinoflagellate cysts appear to be stable for extended periods of time at 8°C. This immediately suggests a cost-effective method for preserving different cultures. In particular, should it become possible to transform dinoflagellate cells, cells with interesting properties can thus be maintained almost indefinitely or shipped world-wide. It would be of interest to determine if encysted cells could be frozen at -80°C to facilitate storage of extensive culture collections.

Figure 5.1. Cyst morphology differs from that of motile cells

(A) Cysts viewed by light microscopy assume a spherical shape and have generally emerged from the theca that normally surrounds motile cells. (B) Scanning electron microscopy of the cysts reveal a generally smooth surface, unlike the discarded theca. Fluorescence microscopy shows reduced numbers of structures with chlorophyll fluorescence. (C-E) Day phase cell and (F-H) cyst taken either under bright field (C, F), or using fluorescence to visualize chloroplasts (D, G) and scintillons (E, H).

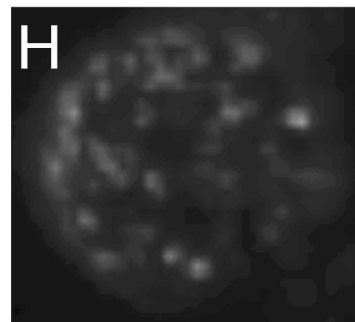
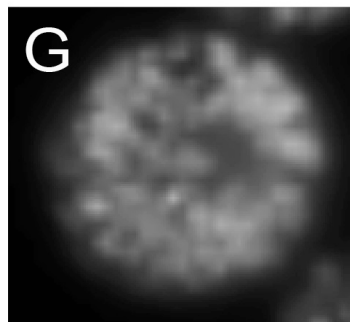
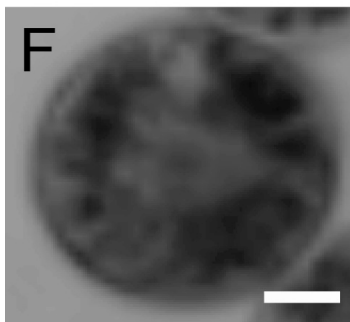
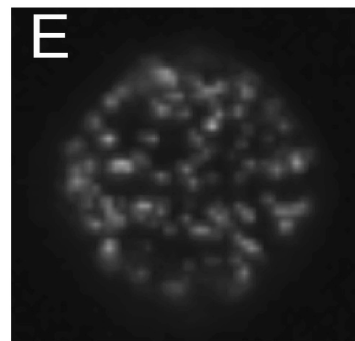
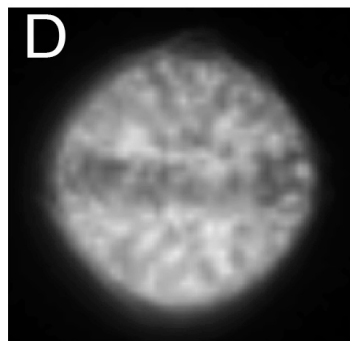
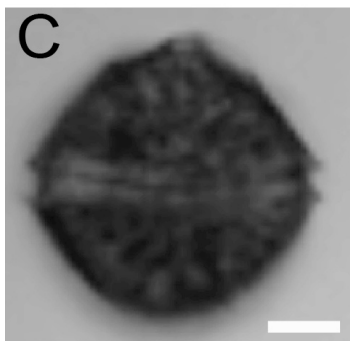
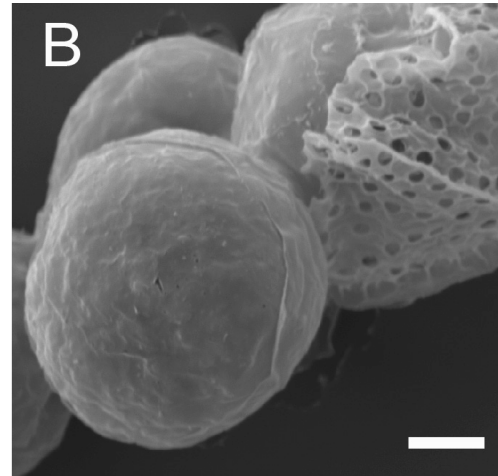
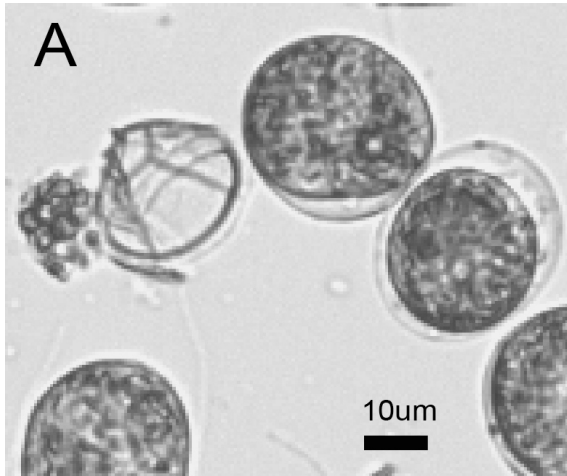


Figure 5.2. Phosphoprotein profiles of cyst and motile cell extracts differ

Protein extracts from cysts and motile cells at two times of day were analyzed by 2D-PAGE. While no differences are observed using the general protein stain Coomassie Blue (left hand panels), cysts show a greatly reduced staining with the phosphoprotein specific stain ProQ diamond (panels at right). Molecular weight markers shown at left are in kDa, and isoelectric points at top show acid and basic ends of the gels.

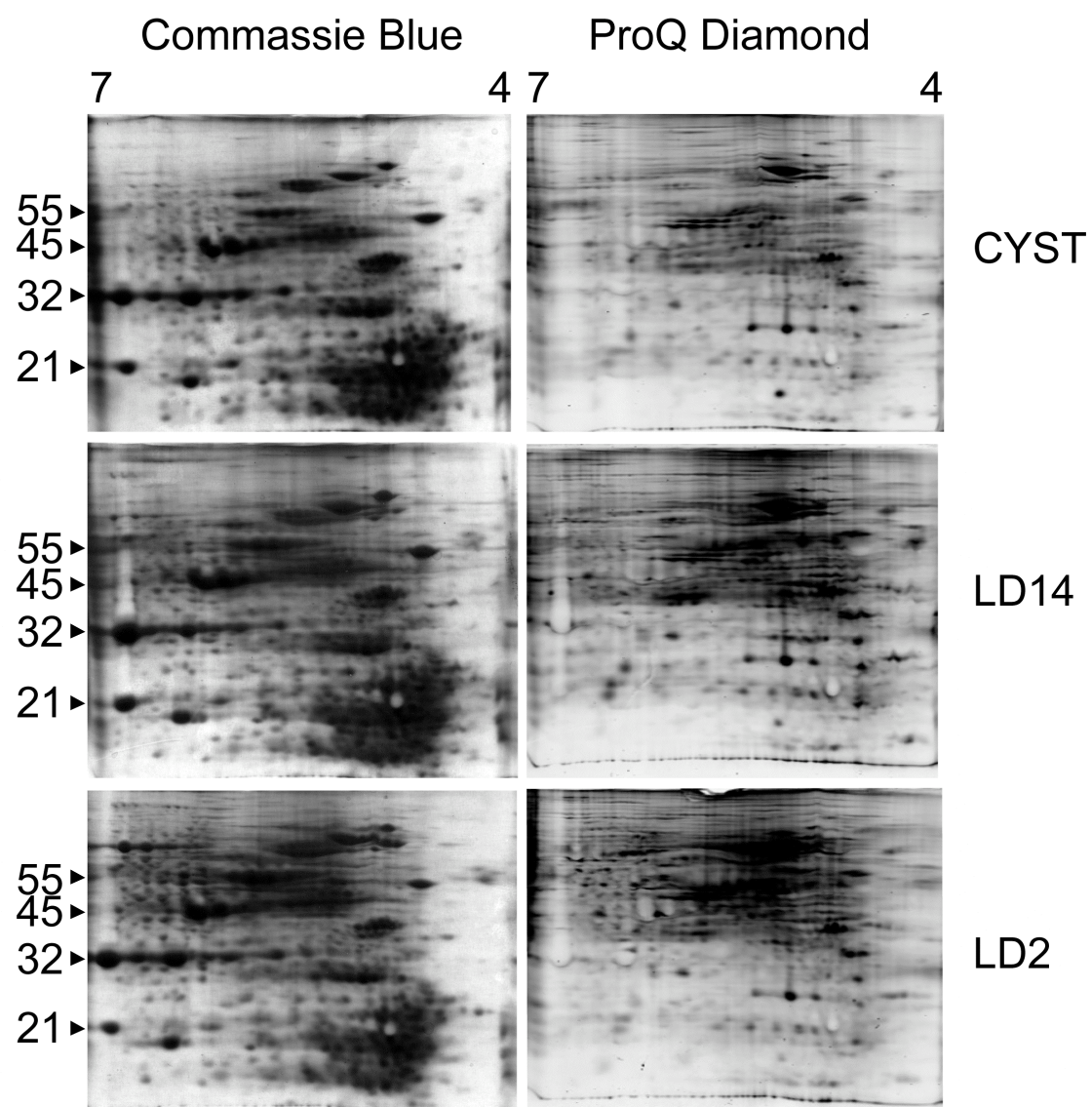


Figure 5.3. Cyst phosphopeptides are generally hypophosphorylated and fall into categories regulating the amount and activity of proteins

(A) A plot of the phosphopeptide intensity ratios between cysts and motile cells that are either greater than the maximum value found in either ZT2 or ZT14 (upper panel) or lower than the minimum value found in either ZT2 or ZT14 (lower panel) shows cysts have a greater number of hypo-phosphorylated peptides. (B) The hypo- and hyper-phosphorylated peptides in the cyst extracts were classified separately into gene ontology (GO) categories. The GO categories relating to translational and post-translational regulation are among the most hypo-phosphorylated in cysts.

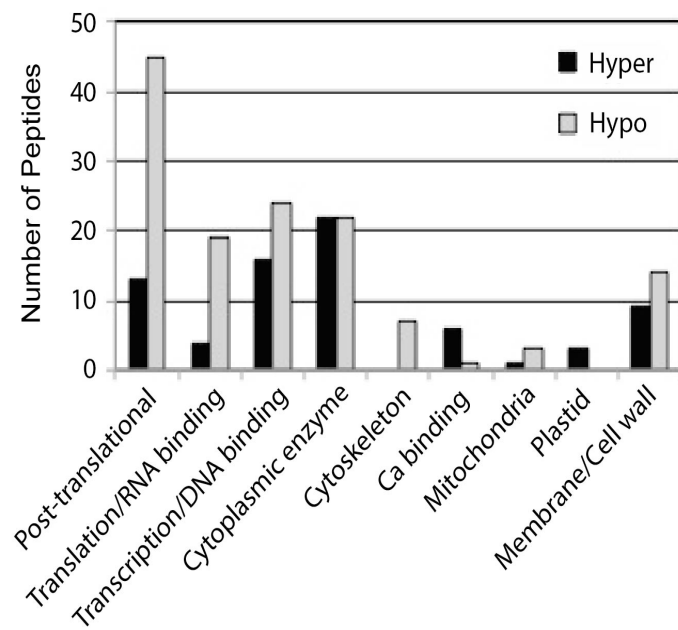
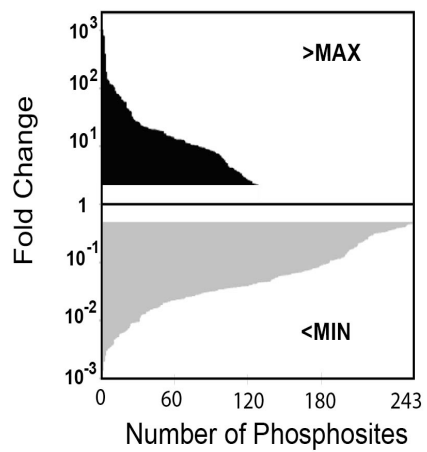


Figure 5.4. Casein Kinase 2 phosphosites are the most hypophosphorylated class

(A) All phosphosites in the ensemble of phosphopeptides identified were first classified into potential kinase target categories based on the peptide motif surrounding the phosphate. The number of phosphosites for each kinase family is shown for both motile cells (ZT2/ZT14) and cysts, with cyst phosphopeptides separated into hypo- and hyper-phosphorylated classes. (B) The ratio between the number of hypo- to hyper-phosphorylated peptides shows that potential CK2 targets in cysts are most affected. (C) Encystment synchronizes the circadian bioluminescence rhythm from two originally out of phase cultures. Bioluminescence of cell cultures placed into constant darkness at the end of the night (upper left) or the end of the day (lower left) shows that the bioluminescence rhythm is out of phase. After an eight-hour exposure to low temperature, the the circadian rhythm of light production has the same phase, independent of the phase prior to encystment. (D) Levels of LBP and PCP were assessed by Western blots using proteins extracted from motile cells at either ZT6 or ZT18 (lanes 1 and 3) or from cells taken at ZT6 and ZT18 but left at 8°C for 12 hours before protein extraction (lanes 2 and 4).

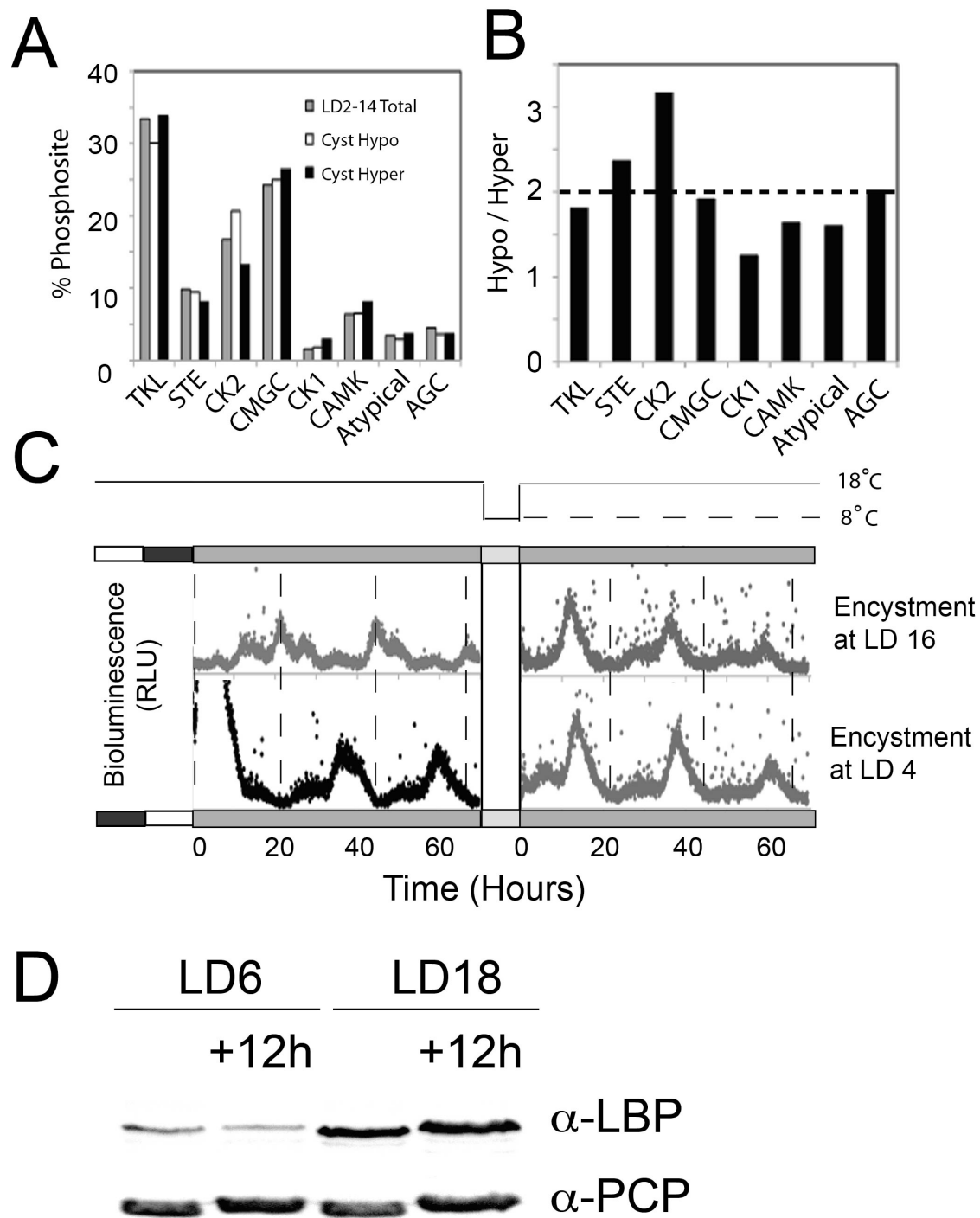
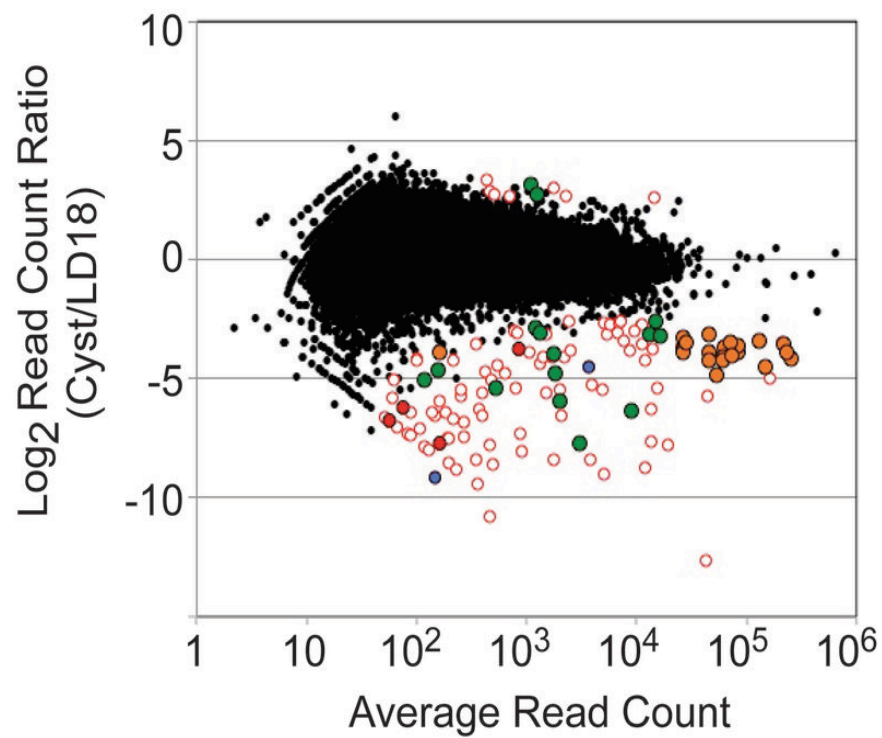
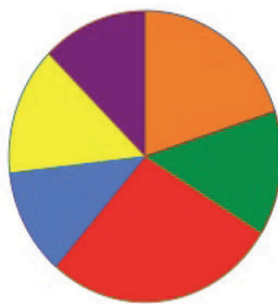


Figure 5.5. RNA-Seq of cyst extracts reveals most RNAs with altered levels have decreased abundance

(A) Read counts from *Lingulodinium* cysts compared to motile cells (harvested at ZT18) using DegSeq. Only 172 RNAs differ significantly in cysts (colored circles). Most of the RNAs with significantly different levels could not be identified (open circles). (B) GO analysis of all RNAs statistically different in cysts compared to motile cells. Plastid-encoded RNA as well as RNAs encoding membrane proteins appear preferentially affected. Color codes correspond to points colored in A.



Transcriptome



Altered in Cyst



- Plastid
- Plasma membrane
- Nucleus
- Cytoplasm
- Mitochondrion
- Secretory pathway

Supplementary Figure Legends

Figure 5.S1. Western blot analysis of three nuclear encoded proteins show no significant decrease in cyst extracts.

Protein extracts prepared from cysts and motile cells taken either during the early day (ZT2) or the early night (ZT14) show no difference in the amount of Luciferin Binding Protein (LBP), Rubisco, or Peridinin Chlorophyll a-Protein (PCP). Extracts were electrophoresed on SDS-PAGE, transferred to nitrocellulose, and probed simultaneously with antibodies recognizing the three different proteins. Antibody reaction was visualized by chemiluminescence. No difference is seen for LBP at these times as ZT14 represents the time of protein synthesis and ZT2 the time of degradation.

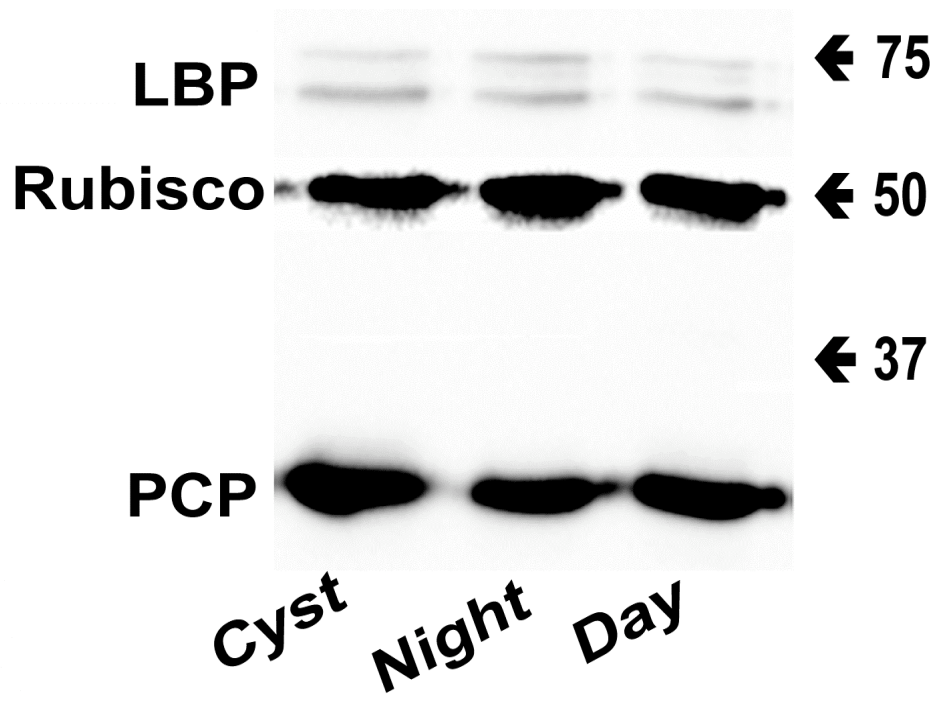


Figure 5.S2. Comparison of phosphoprotein enrichment fraction from cyst and motile cells

The figure shows the comparison of GO-category analysis for the cell component of the 2100 proteins obtained after phosphoprotein enrichment using cyst protein samples with the 3007 proteins obtained from motile (ZT2 and ZT14) cells. The greatest variation between cysts and motile cells is observed in the ribosomal, cytoplasmic and nuclear protein categories, while only the cytoskeleton protein category appears preferentially phosphorylated in cysts.

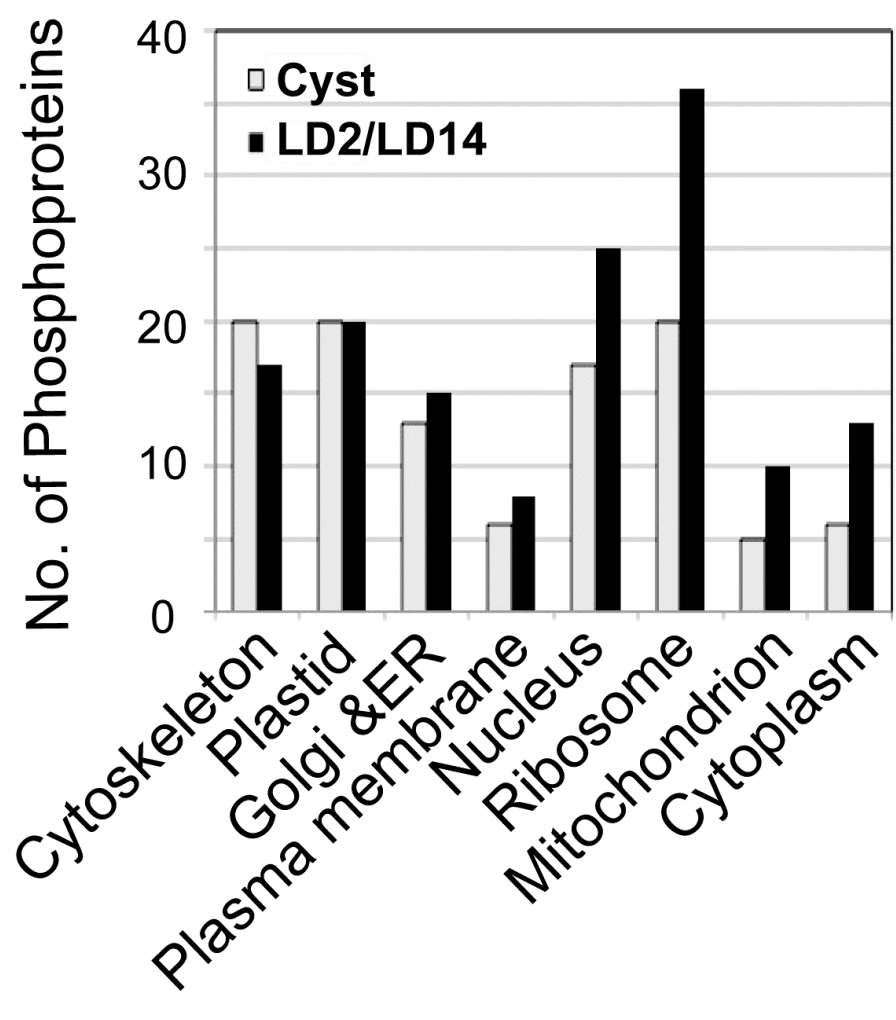
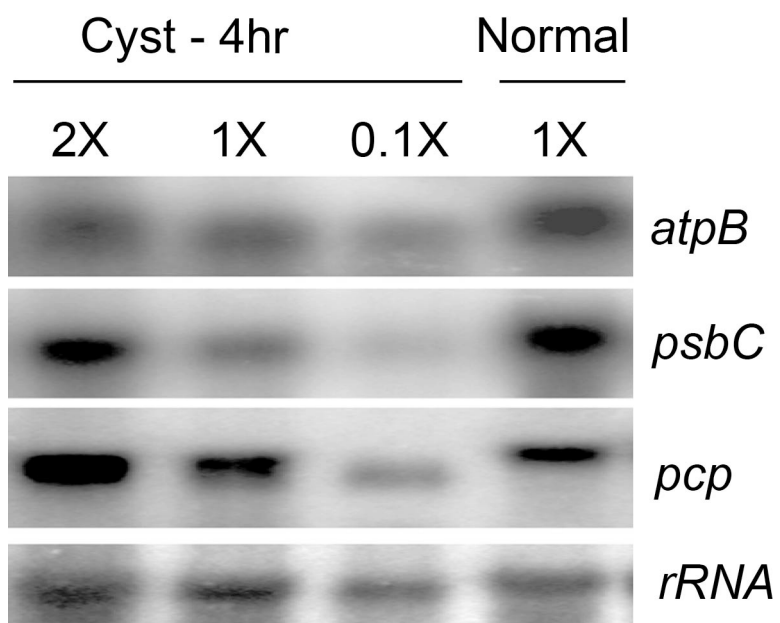


Figure 5.S3. Northern blot analyses of two plastid-encoded and two nuclear-encoded RNAs confirms a decrease in plastid RNAs

RNA was extracted from cysts and motile cells taken at ZT6, electrophoresed through an agarose gel and transferred to a nitrocellulose membrane. The membrane was hybridized sequentially with probes for the plastid encoded *atpB* and *psbC* gene products as well as the nuclear encoded PCP and rRNA gene products. The radiolabeled probes bound to the nitrocellulose were visualized by exposure to a PhosphorImager screen.



Supplementary Tables

Table 5.ST1. Cyst hyperphosphorylated peptides

| Cyst | LD2 | LD14 | UniProt ID | Peptide Sequence | best blast hit description | Peptide Score |
|----------|----------|----------|--------------|---------------------------------|--|---------------|
| 9.97E+06 | 1.00E+04 | 1.00E+04 | GABP01078063 | LGITGLR | Phenylalanine tRNA synthetase | 35.52 |
| 7.98E+06 | 2.00E+06 | 9.00E+04 | GABP01064394 | SPAEGAASIPSRAPAALAK | Protein kinase domain containing protein | 27.23 |
| 7.71E+06 | 1.00E+04 | 1.00E+04 | GABP01020596 | EPSDLNGLRVVATVPAR | Unknown | 21.07 |
| 6.81E+06 | 1.95E+06 | 7.00E+05 | GABP01055210 | KDSANLVCAVMGQMDPGWGGLCQMGMLSK | Type I polyketide synthase | 19.67 |
| 3.75E+06 | 1.00E+04 | 1.00E+04 | GABP01040934 | FQHQERGNTSVCSPLGIAR | Unknown | 16.39 |
| 3.57E+06 | 1.00E+04 | 6.52E+05 | GABP01032554 | EQHVQSVLINPNIAITVQTSK | Pyrimidine biosynthesis protein | 18.5 |
| 2.37E+06 | 1.00E+06 | 5.00E+05 | JO712444 | HGLEGPWHRALLLTTPAR | hypothetical protein Hsro_1709 inosine-5'-monophosphate dehydrogenase 2-like | 23.27 |
| 2.08E+06 | 2.39E+05 | 3.54E+05 | JO707869 | RAGAAPRPTR | Unknown | 35.59 |
| 1.83E+06 | 1.00E+04 | 1.00E+04 | GABP01032701 | KASPLPR | Unknown | 40.58 |
| 1.46E+06 | 1.00E+04 | 4.52E+05 | GABP01107775 | SPAVVTTTGGSPVTMR | Unknown | 68 |
| 1.36E+06 | 1.00E+04 | 1.00E+04 | JO705455 | KLSEVFNESK | glycerol-3-phosphate dehydrogenase | 48.91 |
| 1.32E+06 | 4.48E+05 | 1.70E+05 | GABP01071379 | GVLPLVTQSFVGTDSVIK | Pyruvate kinase | 56.5 |
| 1.28E+06 | 1.00E+04 | 1.00E+04 | GABP01026782 | LQKELLVTAIDLVNADSK | PCNA | 18.55 |
| 1.15E+06 | 1.00E+04 | 1.00E+04 | JO703814 | DTPFVFTLGKQVIGWDLGLVTMR | predicted protein | 22.77 |
| 1.11E+06 | 1.00E+04 | 1.00E+04 | GABP01023199 | TQSGRAAGLSTLLITIVR | GTP binding protein | 19.49 |
| 1.10E+06 | 4.00E+05 | 1.00E+04 | GABP01104632 | LQTRAPVTR | T complex protein 1 subunit | 27.76 |
| 1.09E+06 | 1.00E+04 | 1.00E+04 | GABP01009978 | TGTMVMGGRAAFAANVPGMTGPVSVGR | Unknown | 17.11 |
| 1.02E+06 | 1.00E+04 | 5.41E+04 | JO744440 | HISFSALSIMAEAAEK | Soluble starch synthase 1, chloroplastic/amyloplastic | 88.2 |
| 1.00E+06 | 1.00E+04 | 1.00E+04 | GABP01103272 | HLQLSVEIPVK | calcium dependent protein kinase 7 uncharacterized protein | 21.68 |
| 9.50E+05 | 3.00E+05 | 7.00E+04 | GABP01021399 | APAAGPPGATRPGR | LOC100273704 | 23.62 |
| 8.50E+05 | 1.00E+04 | 1.00E+04 | GABP01012378 | FMAGALATSPAAASR | Unknown | 19 |
| 8.13E+05 | 1.00E+05 | 1.00E+04 | GABP01054309 | LCARSCACQALVSEACAAGAAFFLAGAGVAR | 5'-3' exoribonuclease, putative | 16.19 |
| 7.78E+05 | 1.00E+04 | 1.00E+04 | JO711628 | EEEEFQGSLSRLLELR | zinc finger protein 830 | 15.2 |
| 7.66E+05 | 1.00E+04 | 1.00E+04 | JO732244 | IEISNLNRQFLFR | ubiquitin-activating enzyme e1, putative | 23.09 |
| 7.20E+05 | 3.00E+05 | 1.00E+04 | GABP01040147 | LTIPLLLGTR | EF1 alpha like protein | 23.22 |
| 7.17E+05 | 1.00E+04 | 1.00E+04 | GABP01047998 | TQAEITLETLR | Unknown | 39.37 |
| 7.14E+05 | 1.00E+04 | 3.81E+04 | GABP01043892 | TCFTTRIISAMFNTVTK | UDP glucose dehydrogenase probable ubiquitin-specific processing protease 21 | 17.34 |
| 6.90E+05 | 1.22E+05 | 1.67E+05 | GABP01017846 | ICNQLIPSHLQALLMR | Unknown | 18.49 |
| 6.89E+05 | 1.00E+04 | 1.00E+04 | GABP01017581 | EGQAQELMKAGSIR | Unknown | 21.88 |
| 6.80E+05 | 2.00E+05 | 5.00E+04 | GABP01031902 | SSAAPQSSPTAANGAAK | Unknown | 56.05 |
| 6.07E+05 | 1.00E+04 | 1.00E+04 | GABP01081438 | SQAHQPAWLPSLWSWPYSGTVSVRASWK | Unknown | 22.54 |
| 5.89E+05 | 1.00E+04 | 1.00E+04 | GABP01113640 | ELSQVELDPWLK | P-type ATPase | 18.41 |
| 5.58E+05 | 1.00E+04 | 1.00E+04 | JO762256 | KLMTGMR | malate:quinone oxidoreductase | 29.68 |
| 5.57E+05 | 1.00E+04 | 1.00E+04 | GABP01086251 | LATGSAGSR | Treacle superfamily protein | 17.01 |
| 4.76E+05 | 1.00E+04 | 9.33E+04 | GABP01060877 | ASVSAEAFGDWNKR | cAMP dependent PK regulatory | 58.05 |

| | | | | | subunit | |
|----------|----------|----------|--------------|-------------------------------------|--|--------|
| 4.37E+05 | 1.00E+04 | 1.00E+04 | GABP01026003 | WDRLELSDEETFHPNLDK | cdc37 domain containing protein | 54.63 |
| 4.19E+05 | 1.00E+04 | 1.00E+04 | GABP01107773 | TVVQQRPPSSPSTLSPVGSAPAPASADAATVAR | Unknown | 74.07 |
| 3.94E+05 | 1.00E+04 | 1.00E+04 | GABP01018625 | VSGGAASPGNSR | hypothetical protein HCAG_09240 | 21.2 |
| 3.87E+05 | 1.00E+04 | 1.00E+04 | JO745057 | LCVTITVSSPSFSPHSR | Unknown | 23.26 |
| 3.84E+05 | 9.36E+04 | 6.89E+04 | GABP01073537 | RMACSYCFLNASAWAFHLCTSSIGACALTSSSTR | Unknown | 20.26 |
| 3.35E+05 | 1.00E+05 | 1.00E+04 | GABP01113382 | TGVAALEAPLYQASASTR | C2 calcium dependent | 38.82 |
| 3.33E+05 | 1.00E+05 | 1.27E+05 | JO713775 | ITTWLNSKASK | helicase | 17.94 |
| 3.14E+05 | 1.00E+04 | 1.00E+04 | GABP01113640 | VKNVSWAIGIVVYTGR | P-type ATPase | 16.31 |
| 3.04E+05 | 1.00E+04 | 7.00E+04 | GABP01020434 | DLTGRSHR | Unknown | 17.92 |
| 2.86E+05 | 1.00E+04 | 1.00E+04 | GABP01066635 | HTSAVWDK | Unknown | 24.82 |
| 2.71E+05 | 1.00E+04 | 1.00E+04 | GABP01021946 | RHEAWGAGDAEDDEASEGGDVQK | phosphoglycerate/bisphosphoglycerate mutase | 96.1 |
| 2.59E+05 | 1.00E+04 | 1.00E+04 | JO738779 | YRMVSSSWVAVGEMSSSSYQAVSR | small GTP-binding protein | 16.9 |
| 2.57E+05 | 1.00E+04 | 1.00E+04 | GABP01021814 | SQSTGHLPPKQSMPGASGQLWQAK | histone acetyltransferase type B subunit 2,related | 19.05 |
| 2.51E+05 | 1.00E+04 | 1.00E+04 | JO741393 | TPVRPLEGGNGGTTSAVAPR | Peptidoglycan binding domain containing protein | 15.37 |
| 2.39E+05 | 1.00E+04 | 1.00E+04 | GABP01074928 | NLRTGGIAPVDSGTQLVLRPGLECVR | Unknown | 17.81 |
| 2.38E+05 | 1.00E+04 | 8.40E+04 | GABP01110879 | DASSLQTLISGLQMLISGLPCGRTCLQAP | Dual specificity phosphatase | 32.02 |
| 2.25E+05 | 5.00E+04 | 1.00E+04 | GABP01062726 | SSRPPTTGPR | AFG1 family ATPase | 16.07 |
| 2.24E+05 | 1.00E+04 | 1.00E+04 | GABP01022104 | GQVFDLPIATTFREIVQAGSPSAK | DNA topoisomerase 3-beta-1 | 20.18 |
| 2.17E+05 | 1.00E+04 | 1.00E+04 | JO703895 | QEEEEEDDDSDVDNAVHAEEK | ABC transporter, putative | 82.47 |
| 2.10E+05 | 1.00E+04 | 1.00E+04 | GABP01072521 | KQASSLQER | Threonyl tRNA synthetase | 50.99 |
| 2.09E+05 | 9.88E+04 | 9.42E+04 | GABP01007146 | ALNIVLASGLAK | Unknown | 21.2 |
| 2.08E+05 | 1.00E+04 | 1.00E+04 | GABP01019685 | SLNKLINVIK | Unknown | 18.45 |
| 2.05E+05 | 1.00E+04 | 8.39E+04 | JO753235 | VLVSQISEPAVKYSLWMSTMSSGLLK | vitamin B12 dependent methionine synthase | 18.6 |
| 2.05E+05 | 1.00E+04 | 1.00E+04 | GABP01010336 | AEAVDLEAGSPVQAR | acyltransferase 3 | 38.21 |
| 2.02E+05 | 1.00E+04 | 1.00E+04 | GABP01017472 | SSNVSSNFIGLEPSSLNVFMK | Flavonol 4'-sulfotransferase | 31.66 |
| 1.93E+05 | 1.00E+04 | 1.00E+04 | JO708022 | GYSPPPAK | hypothetical protein Pmar_PMAR024533 | 24.82 |
| 1.92E+05 | 1.00E+04 | 1.00E+04 | JO731992 | HAMGPHQVPQAYGEVSPAR | ALVEOLIN1 | 53.45 |
| 1.89E+05 | 1.00E+04 | 1.00E+04 | GABP01020596 | KATTGSIDLPMQAQSK | Unknown | 102.07 |
| 1.88E+05 | 1.00E+04 | 1.00E+04 | GABP01031041 | FFNLSNPNEAR | type I polyketide synthase-like protein | 24.58 |
| 1.87E+05 | 1.00E+04 | 1.00E+04 | GABP01114261 | QVLEVSYIALQGAGWTKK | type I polyketide synthase-like protein KB5361 | 16.51 |
| 1.85E+05 | 1.00E+04 | 1.00E+04 | GABP01029672 | MYSRTNWANVSLVMVLAPSSQR | Calcium binding EF-hand conytaining protein | 24.74 |
| 1.83E+05 | 1.00E+04 | 1.00E+04 | JO720314 | KVPVEGGLGDISGGDS | PREDICTED: similar to calcium-binding protein Calnexin | 34.81 |
| 1.80E+05 | 1.00E+04 | 1.00E+04 | JO708230 | LSVEVSPSNQAQR | hypothetical protein BRAFLDRAFT_73148 | 28.92 |
| 1.78E+05 | 1.00E+04 | 1.00E+04 | GABP01030162 | AAVGDFLEPDGDEGDPSPR | Unknown | 41.81 |
| 1.76E+05 | 1.00E+04 | 1.00E+04 | JO703895 | AQSEVDGEGEELPATEK | ABC transporter, putative | 104.34 |
| 1.76E+05 | 1.00E+04 | 1.00E+04 | GABP01020476 | KPEAAEPAAAAVPGAGPPGSPAAAGAVAAPLAAGR | Unknown | 54.8 |
| 1.76E+05 | 8.00E+04 | 1.00E+04 | GABP01025704 | MGSLSLVLEFAHLSASLSLRR | protein with transferase activity | 16.62 |
| 1.59E+05 | 1.00E+04 | 1.00E+04 | GABP01082104 | RVLPLFEASMAEFTHR | Cleavage and polyadenylation specificity factor 73 kDa subunit | 17.98 |

| | | | | | | |
|----------|----------|----------|--------------|--------------------------------|---|--------|
| 1.58E+05 | 1.00E+04 | 1.00E+04 | GABP01000126 | GPAGVSSILQVFMK | Unknown | 18.09 |
| 1.50E+05 | 4.00E+04 | 1.00E+04 | JO734365 | DTVRGVSGQATGSEVR | Unknown | 45.47 |
| 1.45E+05 | 1.00E+04 | 1.00E+04 | GABP01080861 | KGSEGGELK | myotubularin-like | 18.56 |
| 1.44E+05 | 1.00E+04 | 1.00E+04 | GABP01012335 | QLQIVRSAR | Unknown | 27.98 |
| 1.44E+05 | 1.00E+04 | 1.00E+04 | JO759300 | SFEDEPTAYSVESQQAQTDGSPASQSHEK | LuxR-family transcriptional regulator | 20.24 |
| 1.42E+05 | 1.00E+04 | 1.00E+04 | GABP01103520 | STGAPRPDALPGDER | Unknown | 22.53 |
| 1.39E+05 | 1.00E+04 | 1.00E+04 | JO714787 | TASRAGASR | hypothetical protein SORBIDRAFT_03g004420 | 31.01 |
| 1.38E+05 | 1.00E+04 | 1.00E+04 | GABP01039948 | IRDPPSALR | Unknown | 31.06 |
| 1.31E+05 | 1.00E+04 | 4.42E+04 | GABP01023642 | NRGDDISEGGNR | Unknown | 29.98 |
| 1.30E+05 | 1.00E+04 | 1.00E+04 | JO764106 | AMMCSSLVTLSIGNNQMGDVEYGK | ribulose 1,5-bisphosphate carboxylase oxygenase form II | 32.07 |
| 1.29E+05 | 1.00E+04 | 1.00E+04 | GABP01058502 | SSPAAEQFEQQGSPAR | Calcium ion binding protein | 75.53 |
| 1.26E+05 | 1.00E+04 | 1.00E+04 | JO708551 | GAGSSASPVPETLSMTGLQKR | Glutamate 5-kinase, putative | 15.85 |
| 1.26E+05 | 1.00E+04 | 1.00E+04 | GABP01088064 | SPAPRPWASGGSRPRTR | Unknown | 15.12 |
| 1.26E+05 | 1.00E+04 | 1.00E+04 | GABP01040969 | ESESGSRPQEEVSQDLPADRSDD | RRM domain containing protein | 42.03 |
| 1.25E+05 | 1.00E+04 | 1.00E+04 | GABP01106179 | HSSRGPQAVVAALALR | proteophosphoglycan 5 | 25.97 |
| 1.24E+05 | 1.00E+04 | 1.00E+04 | GABP01098060 | AVPADAVIEHSPPEAR | Unknown | 56.21 |
| 1.24E+05 | 1.00E+04 | 1.00E+04 | JO712445 | SPCCSAAGSSSATTLDWPGSSPPTPPGEAR | Unknown | 21.84 |
| 1.22E+05 | 5.00E+04 | 1.00E+04 | GABP01093678 | RPPPSPPR | 2OG-Fe(II) oxidoreductase like- protein | 62.41 |
| 1.17E+05 | 1.00E+04 | 1.00E+04 | JO715203 | SADVAEYTGPDGSPQAR | polymorphic outer membrane protein | 54.11 |
| 1.15E+05 | 5.00E+04 | 1.00E+04 | GABP01046571 | RLPGLAGGHLLAAASPSR | Unknown | 38.11 |
| 1.10E+05 | 1.00E+04 | 1.00E+04 | GABP01038652 | ASEETPEGAQEPGLASPHKPSTAK | Calcium ion binding protein | 25.65 |
| 1.09E+05 | 1.00E+04 | 1.00E+04 | GABP01013748 | TAAVASPEQAEGGGDEAER | Unknown | 106.59 |
| 1.08E+05 | 1.00E+04 | 1.00E+04 | GABP01092031 | GTCQANCLSAALSRK | Heat shock DNAJ N terminal | 20.31 |
| 1.07E+05 | 1.00E+04 | 1.00E+04 | GABP01007088 | ELFGTDEETAPPAQVVAATPDK | Leo1p | 17.34 |
| 1.03E+05 | 1.00E+04 | 2.59E+04 | GABP01087772 | ASSWRPPGGASSR | Unknown | 40.19 |
| 1.02E+05 | 1.00E+04 | 1.00E+04 | GABP01107414 | SEYRDNSDSGY | CSD protein | 26.22 |
| 1.01E+05 | 1.00E+04 | 1.00E+04 | GABP01012985 | DLAQAGGGGGGGGGGPTSPTR | calcium-binding protein | 85.83 |
| 1.01E+05 | 1.00E+04 | 4.16E+05 | JO695170 | SSSGLEDLQVEDGDGGSKK | KH domain similar to putative chromatin structure regulator | 67.5 |
| 9.88E+04 | 1.00E+04 | 1.00E+04 | JO757628 | TPAGVPVHSPGGVPVQTPAGVPSTPAGAPR | Unknown | 41.48 |
| 9.86E+04 | 1.00E+04 | 1.00E+04 | GABP01056670 | RGGESDAEEAK | Unknown | 50.55 |
| 9.82E+04 | 1.00E+04 | 1.00E+04 | GABP01099907 | SLPAADTEESGDEDAK | Unknown | 58.15 |
| 9.43E+04 | 1.00E+04 | 1.00E+04 | GABP01041071 | VLSFAPSDGEEAEVAPK | Isomerase activity containing protein | 26.19 |
| 9.34E+04 | 1.00E+04 | 1.00E+04 | GABP01011334 | VSGNAPQSAVRQSAASLTVAAWVSLVCTR | Putative sodium-coupled neutral amino acid transporter 10 | 18.92 |
| 9.26E+04 | 6.00E+04 | 1.00E+04 | JO747619 | KPEDEASDNDGSEAK | Unknown | 16.4 |
| 9.12E+04 | 1.00E+04 | 1.00E+04 | GABP01024922 | EADAGSEEEAEETGR | pinin/sdk/mema domain containing protein | 70.18 |
| 8.83E+04 | 1.00E+04 | 1.00E+04 | GABP01029523 | TRTAPPSTTSR | AAA type ATPase domain containing protein | 17.99 |
| 8.58E+04 | 1.00E+04 | 1.00E+04 | JO727534 | AAAKEEDSDDEEEK | DnaJ/SEC63 protein | 75.22 |
| 8.51E+04 | 3.59E+04 | 3.94E+04 | GABP01004963 | AWPSGTANPSAGPTILSTASPPSK | ubiquitin carboxyl-terminal hydrolase isozyme L5 | 18.6 |
| 8.48E+04 | 1.00E+04 | 1.00E+04 | GABP01082925 | QDSHGFDLIR | Fumarate mitochondrial precursor | 26.54 |
| 8.44E+04 | 1.00E+04 | 1.00E+04 | JO701367 | AASPGLGGEGTPSSAVR | AAA ATPase | 47.92 |

| | | | | | | |
|----------|----------|----------|--------------|----------------------------------|--|-------|
| 7.96E+04 | 1.00E+04 | 1.00E+04 | GABP01054222 | CACFAMATGSWKTAASR | Unknown | 32.32 |
| 7.87E+04 | 1.00E+04 | 1.00E+04 | GABP01072459 | MVCASRFEMPEGMR | Unknown | 26.47 |
| 7.76E+04 | 1.00E+04 | 1.00E+04 | JO740006 | GAAASPGVLLPR | Unknown | 18.9 |
| 7.43E+04 | 1.00E+04 | 1.00E+04 | JO715544 | ASSSAQEGGEEESDEADAAQRPQQAQK | Unknown | 45.15 |
| 7.24E+04 | 1.00E+04 | 1.00E+04 | JO716068 | NSWIMACTFGMRVEPPTSTTSFTCLLSMPLSR | Hsp70 | 15.87 |
| 7.16E+04 | 1.00E+04 | 1.00E+04 | GABP01027075 | FDQIEDSDDEREK | beta-hydroxylase | 35.19 |
| 6.81E+04 | 1.00E+04 | 1.00E+04 | GABP01085587 | AASEEGAEAAK | Unknown | 58.32 |
| 6.56E+04 | 1.00E+04 | 1.00E+04 | JO757628 | TPAGPVPVHSPGGVPVQTPAGPVPSTPAGAPR | similar to putative chromatin structure regulator | 32.95 |
| 6.15E+04 | 2.78E+05 | 2.78E+05 | GABP01054190 | LAPHSAA SPEPSRPRSPEPAER | CSD protein | 18.94 |
| 1.95E+05 | 1.00E+04 | 1.00E+04 | JO721439 | KWADVDEEDEEGFNESPK | eukaryotic translation initiation factor 3 p42 subunit | 93.82 |

Table 5.ST2. Cyst hypophosphorylated peptides

| Cyst | LD2 | LD14 | UniProt ID | Peptide Sequence | best blast hit description | Score |
|----------|----------|----------|--------------|--|--|--------|
| 1.00E+04 | 3.00E+05 | 2.00E+05 | GABP01036833 | AAQEPAAGEDQGEAAP PT TPPAAAGSK | snRNP Gar1 RNA binding | 54.63 |
| 1.00E+04 | 4.00E+05 | 5.00E+05 | J0733358 | AAS A SELLEK | hypothetical protein SORBIDRAFT_06g025060 | 40.22 |
| 1.00E+04 | 4.00E+05 | 5.00E+05 | GABP01025294 | AAVQSTSALAQC P SLR | RNA pol II associated protein | 18.26 |
| 2.00E+05 | 9.00E+05 | 1.00E+06 | GABP01086664 | ADIFDDLPGENWVG S PAMR | Protein kinase | 37.32 |
| 1.00E+04 | 3.00E+05 | 9.00E+04 | J0719969 | AD S DEE S EEEEPPQKK | RNA recognition motif. family protein | 22.42 |
| 1.00E+04 | 4.00E+05 | 3.00E+05 | GABP01076421 | AEAAAGAAAA A SALQ | Unknown | 26.39 |
| 4.00E+05 | 3.00E+06 | 1.00E+06 | GABP01084062 | AEEDPPSKPQED A DDDAKGK | Unknown | 45.16 |
| 1.00E+04 | 4.00E+05 | 1.00E+05 | J0707688 | AEQVELQ S PAR | cell division protein FtsY, putative | 30.39 |
| 1.00E+04 | 1.00E+05 | 2.00E+05 | GABP01000748 | AE S DDVAADVFPAPMETDEEVAR | Serine threonine protein kinase prp4 | 91.53 |
| 1.00E+04 | 4.00E+05 | 8.00E+05 | J0707225 | AFVL S FTQLAGA | Hsp90 | 15.98 |
| 2.00E+05 | 4.00E+05 | 4.00E+05 | GABP01024922 | AGDDG A SDREADAG S EEEEAEETGR | Pinin/SDK/Mema domain containing protein | 70.34 |
| 1.00E+04 | 9.00E+05 | 1.00E+06 | GABP01013062 | AGIEDGDEQDAEAAAHGGDQRP V SR | Unknown | 85.29 |
| 1.00E+04 | 4.00E+05 | 4.00E+05 | J0714961 | AGRAAPP T GPSHAAHAAASK | Unknown | 24.19 |
| 1.00E+04 | 5.00E+05 | 4.00E+05 | GABP01038472 | AG S EEAGAAEPASNVGK | Unknown | 101.93 |
| 1.00E+04 | 1.00E+06 | 2.00E+06 | J0708271 | AG S ITDDVFNMVDR | ALVEOLIN1 [Karlodinium veneticum] | 85.49 |
| 9.00E+04 | 3.00E+05 | 2.00E+05 | GABP01012557 | AKG S DGEDEDEEPEPK | histone deacetylase | 93.02 |
| 3.00E+05 | 1.00E+06 | 1.00E+06 | GABP01104797 | AKKPAPAEED S DEE | RRM domain containing protein | 59.45 |
| 1.00E+04 | 3.00E+05 | 2.00E+05 | GABP01071950 | ALALALPPPPGP A SLPVPTGSGR | Unknown | 17.03 |
| 1.00E+04 | 2.00E+06 | 2.00E+06 | GABP01108400 | ALGMAPGEVSPIEHLLFGPD T PAEELQVVSS LWAP | Unknown | 19.96 |
| 3.00E+05 | 2.00E+06 | 1.00E+06 | GABP01031954 | ALRPEAGSGPG S PSR | Unknown | 23.05 |
| 8.00E+05 | 2.00E+06 | 3.00E+06 | GABP01024922 | AL S PAAAAGGPGQGR | Pinin/SDK/Mema domain containing protein | 73.02 |
| 1.00E+04 | 1.00E+06 | 8.00E+04 | GABP01054884 | AMRLSC S HQALR | Dynein heavy chain | 15.69 |
| 1.00E+04 | 2.00E+05 | 9.00E+04 | GABP01041886 | AMSS S CAACG S ST A T S SS S GTSR | Unknown | 18.82 |
| 1.00E+04 | 3.00E+05 | 4.00E+05 | GABP01049570 | APDDTSMFDRY P ESTEGAPSISQADQEHFE GFGK | cAMPdependent protein kinase | 62.04 |
| 1.00E+04 | 2.00E+06 | 1.00E+06 | GABP01051835 | AP S PAPAAPAPQAR | Unknown | 56.07 |
| 1.00E+04 | 2.00E+05 | 1.00E+05 | GABP01047267 | AQAPKPPGSM P STPAGEVAAAK | Unknown | 46.14 |
| 1.00E+04 | 1.00E+05 | 8.00E+04 | GABP01085385 | AQQAEPQQ A SPEAK | High Mobility group protein | 72.69 |
| 1.00E+04 | 2.00E+06 | 1.00E+06 | GABP01084062 | ARAEEDPPSKPQED A DDDAK | Unknown | 42.7 |
| 1.00E+04 | 8.00E+05 | 4.00E+05 | J0759786 | ARG S QLPL | hypothetical protein Pmar_PMAR019944 | 15.56 |
| 1.00E+04 | 7.00E+05 | 1.00E+07 | GABP01073223 | ARPPPAAS P ASHLR | tetrahydrofolate dehydrogenase/cyclohydrolase | 15.2 |
| 1.00E+05 | 2.00E+05 | 3.00E+05 | J0754053 | ASGFTEDNN S DDEQEYQPDPAEVVEVSHG K | DNAj domain, possible transmembrane domain | 46.51 |
| 1.00E+04 | 5.00E+06 | 2.00E+06 | J0703895 | ATGT E EA S DAEDAR | ABC transporter | 60.58 |
| 1.00E+04 | 4.00E+05 | 7.00E+05 | GABP01026436 | AVEDEEE S DDCEIPESFKKPEAQMGR | cAMP dependent protein kinase regulatory | 127.85 |

| | | | | | | |
|----------|----------|----------|--------------|---|--|--------|
| 3.00E+05 | 3.00E+06 | 1.00E+06 | J0720544 | AVGTPQNAASGGGSATK | NOT2/NOT3/NOT5 domain-containing protein | 125.79 |
| 1.00E+05 | 2.00E+06 | 1.00E+06 | GABP01028776 | AVTEPVPTQQASPDASPTK | ARF-GAP like | 64.39 |
| 1.00E+04 | 4.00E+05 | 3.00E+05 | GABP01029968 | AYDDADTDDEDAKPR | SRP 68kDa | 71.68 |
| 6.00E+04 | 3.00E+05 | 4.00E+05 | GABP01054309 | CLRPSATTSMTAPPMTR | 5'-3' exoribonuclease | 16.97 |
| 1.00E+04 | 2.00E+06 | 3.00E+06 | GABP01051532 | CLYTRLAVQATQK | Unknown | 20.19 |
| 1.00E+04 | 2.00E+05 | 2.00E+05 | GABP01061067 | DAPPIGPGRHAPARPEVTGTGSVQAAR | Unknown | 26.62 |
| 1.00E+04 | 1.00E+05 | 4.00E+04 | GABP01038927 | DEDSSEDEQEPK | Formin binding | 56.44 |
| 1.00E+04 | 3.00E+05 | 4.00E+05 | GABP01007088 | DLFGSEDEGPEIDER | Rna polymerase associated protein Leo1 | 57.57 |
| 1.00E+04 | 9.00E+04 | 2.00E+05 | GABP01022683 | DPLLLGAGSAMGPVKSEPIA | Monovalent cation proton antiporter | 46.08 |
| 3.00E+05 | 9.00E+05 | 8.00E+05 | GABP01049965 | DRAPTPEASEEANVNR | Pumilio | 35.49 |
| 1.00E+04 | 3.00E+05 | 3.00E+05 | GABP01046617 | EEEDPSSEEEEEAKEEGAAEK | Unknown | 42.84 |
| 1.00E+04 | 3.00E+05 | 4.00E+05 | GABP01000248 | EEEEDEGASDEDEVEQK | Unknown | 55.36 |
| 1.00E+05 | 1.00E+06 | 8.00E+05 | J0702217 | EEEEKESEEEPPAPNAELTK | hypothetical protein Pmar_PMAR024059 | 63.91 |
| 1.00E+04 | 1.00E+05 | 2.00E+05 | GABP01036280 | EESPPQPSGHGDPYEGFFGSLPPDEFLLPQEMD LR | Unknown | 55.75 |
| 9.00E+04 | 8.00E+05 | 2.00E+05 | GABP01091378 | EGDVITCLLDRENQTISYCK | putative ATP-dependent RNA helicase DDX family protein | 20.7 |
| 1.00E+04 | 2.00E+05 | 4.00E+04 | GABP01066887 | EGEGAQGAEASPAASK | Armadillo like domain containing | 55.82 |
| 1.00E+04 | 1.00E+05 | 3.00E+05 | GABP01007088 | EIFGDISDDEEPEKVEDVILR | Rna polymerase associated protein Leo1 | 76.7 |
| 1.00E+04 | 2.00E+05 | 1.00E+05 | J0743508 | EKEVTDSEDEEECK | heat shock protein 90 1 | 40.88 |
| 8.00E+05 | 3.00E+06 | 3.00E+06 | J0744240 | EPLLLVGSFNSWSVEKAR | Unknown | 19.14 |
| 1.00E+05 | 8.00E+05 | 3.00E+05 | GABP01042912 | EPVPAPETTTANEPSPK | RNA pol II associated protein 3 | 40.27 |
| 1.00E+04 | 2.00E+05 | 2.00E+05 | GABP01081001 | EQCQGCNDNMTLMVVDLQAGGAGCPAASGG LQAASPGTASVNGAAQ | Protein Phosphatase | 16.94 |
| 1.00E+04 | 3.00E+06 | 1.00E+06 | GABP01070797 | FAASMTPGPPPVQTLVYGSFPR | Rossmann fold nucleotide-binding protein | 19.88 |
| 1.00E+04 | 4.00E+05 | 3.00E+05 | GABP01019831 | FDEIEDSDEKTQEK | TPR domain protein | 37.03 |
| 1.00E+04 | 7.00E+05 | 4.00E+05 | GABP01068216 | FDHIEDSDETPAAKPVPK | Unknown | 82.44 |
| 1.00E+04 | 1.00E+06 | 6.00E+05 | GABP01051606 | FDNIEDSDEKPK | Unknown | 38.41 |
| 1.00E+04 | 1.00E+05 | 5.00E+05 | J0758863 | FGATRSILALR | Nucleolar protein Nop56 | 17.22 |
| 1.00E+04 | 1.00E+05 | 3.00E+05 | GABP01114472 | FHSVQDETFLGDTK | Endoplasmic precursor,(HSP 90 like) | 69.03 |
| 1.00E+06 | 5.00E+06 | 4.00E+06 | GABP01042453 | FLQLLAVLSLCQPR | Unknown | 24.46 |
| 1.00E+04 | 5.00E+04 | 8.00E+04 | GABP01029969 | FQITALDEPMELTSACASWR | Unknown | 22.34 |
| 9.00E+04 | 3.00E+05 | 3.00E+05 | GABP01018703 | FTTGPEEAEPGSEGESEAR | cGMP-dependent protein kinase | 53.32 |
| 1.00E+04 | 4.00E+05 | 1.00E+05 | GABP01052175 | GAAESSRVSLPAPAVAR | cap-gly domain containing linker protein partial | 24.74 |
| 1.00E+04 | 2.00E+05 | 3.00E+05 | GABP01024922 | GAAGEAAEAADAEAAASPGEEGK | Pinin/SDK/Mema domain containing protein | 134.06 |
| 1.00E+04 | 4.00E+05 | 3.00E+05 | GABP01086521 | GADAPAVAAASPPVPVLPQGR | LsmAD domain containing protein | 48.18 |

| | | | | | | |
|----------|----------|----------|--------------|---|---|--------|
| 2.00E+05 | 1.00E+06 | 5.00E+05 | GABP01024074 | GAEVDDSPPR | Unknown | 29.94 |
| 1.00E+04 | 2.00E+05 | 7.00E+05 | GABP01031190 | GAPGSVTSQHRATIAAMELNCSR | DEAD-box RNA/DNA helicase | 16.8 |
| 1.00E+04 | 2.00E+05 | 2.00E+05 | J0757628 | GFAPQTPAGLGQMPATPAPHPATPAGVA PMSPAR | PREDICTED: similar to putative chromatin structure regulator | 59.97 |
| 1.00E+04 | 4.00E+05 | 4.00E+05 | GABP01079143 | GGFRMATIGGSR | Unknown | 29.78 |
| 1.00E+04 | 7.00E+06 | 8.00E+06 | GABP01033861 | GGRTCHAAIHAR | Phosphoenol pyruvate synthase | 39.26 |
| 1.00E+04 | 1.00E+06 | 4.00E+05 | J0709548 | GHAGPSAVAQQHTPPSHACLHLLAAR | Unknown | 16.92 |
| 1.00E+04 | 2.00E+05 | 1.00E+05 | J0709540 | GICGGNMPYIEGSAKSPAGTVLTLR | hypothetical protein BRAFLDRAFT_286861 | 15.3 |
| 1.00E+04 | 6.00E+05 | 8.00E+05 | GABP01061001 | GKSTESLTDEEQAK | 2OG-Fe(II) oxygenase superfamily protein | 92.68 |
| 1.00E+04 | 5.00E+05 | 3.00E+05 | J0734709 | GLAAVLGSVALAVGTGPR SAR | predicted protein | 16.37 |
| 1.00E+04 | 5.00E+05 | 3.00E+05 | J0717527 | GLAKDEES EEEQER | hypothetical protein CHLNCDRAFT_136322 | 105.34 |
| 1.00E+04 | 5.00E+05 | 3.00E+05 | J0699741 | GLELDDVILRTPHQLTR | predicted protein | 23.46 |
| 2.00E+05 | 3.00E+06 | 8.00E+05 | J0748411 | GLLAGQGGGLLVLLSGRDR | Unknown | 18.49 |
| 1.00E+04 | 3.00E+05 | 3.00E+05 | GABP01104468 | GLLPSSGEGDSPVDSAR | Unknown | 30.66 |
| 1.00E+04 | 9.00E+05 | 6.00E+05 | GABP01092558 | GPASDLELADSSGR | DNAJ heat shock N-terminal domain- containing protein | 16.78 |
| 1.00E+04 | 3.00E+05 | 2.00E+05 | GABP01078304 | GSGTRSTSRPAASLGGTIAGSFCWT | chalcone and stilbene synthase domain-containing protein | 16.37 |
| 1.00E+04 | 1.00E+05 | 2.00E+05 | J0695062 | GVLPLVTQSFVGTDSVIK | Pyruvate kinase | 76.57 |
| 3.00E+05 | 1.00E+06 | 9.00E+05 | GABP01071379 | GVLPLVTQSFVGTDSVIK | Pyruvate kinase, | 105.27 |
| 8.00E+04 | 2.00E+05 | 3.00E+05 | J0750329 | GYGPEEDLHGSPD SWVEPEALWEFADDA ESR | carbonic anhydrase 2 | 98.46 |
| 1.00E+05 | 2.00E+06 | 1.00E+06 | J0748908 | HLQSADAEPDGPVR | calcium-dependent protein kinase | 64.05 |
| 1.00E+04 | 3.00E+05 | 3.00E+05 | GABP01040066 | HMQESDEESPTGLEK | Unknown | 65.72 |
| 1.00E+04 | 3.00E+05 | 5.00E+05 | GABP01008812 | IAIIFTSIVIR | Unknown | 38.14 |
| 1.00E+04 | 1.00E+06 | 2.00E+06 | J0752062 | ICGSKPISNIR | DEAD-box ATP-dependent RNA helicase | 23.4 |
| 1.00E+04 | 1.00E+05 | 1.00E+05 | J0734675 | IEDLSAQAAQASAAQQFTQQGLEAAVEAAET SAEAPK | Nascent polypeptide-associated complex subunit alpha | 43.8 |
| 1.00E+04 | 1.00E+05 | 3.00E+05 | GABP01036009 | ISAVIESVPDKSPR | lacto-N-neotetraose biosynthesis glycosyl transferase LgtA | 57.47 |
| 1.00E+04 | 2.00E+05 | 2.00E+05 | J0749092 | ISSGGGALCPSAAWLAPR | acetyl-CoA carboxylase | 16.57 |
| 1.00E+04 | 6.00E+04 | 7.00E+04 | GABP01000274 | ISVTSAVTFHPAGGASLPVL | Unknown | 21.77 |
| 1.00E+04 | 3.00E+05 | 2.00E+05 | GABP01086264 | ITEINMLKSLK | Protein kinase domain containing protein | 23.3 |
| 6.00E+04 | 1.00E+05 | 1.00E+05 | J0731992 | IVEVPTVCTQEVVKAVPK | ALVEOLIN1 [Karlodinium veneficum] | 23.49 |
| 1.00E+04 | 4.00E+05 | 1.00E+05 | J0719969 | KAAVEDSDDSDDEPAPK | RNA recognition motif. family protein | 71.13 |
| 2.00E+05 | 2.00E+06 | 1.00E+06 | J0758863 | KASAADAEAEPPAEAPK | Nucleolar protein Nop56 | 116.44 |
| 1.00E+04 | 3.00E+05 | 2.00E+05 | GABP01062475 | KESAEAAAAEAEAPPPEK | Nucleolar protein | 102.47 |
| 1.00E+05 | 3.00E+05 | 2.00E+05 | GABP01028630 | KPAPPVGRSLR | C2 domain containing protein | 23.71 |
| 1.00E+04 | 2.00E+06 | 7.00E+05 | GABP01024454 | KPGEEDGSPHSLQEYSK | Unknown | 70.79 |
| 1.00E+04 | 3.00E+05 | 1.00E+05 | J0709869 | KPKPLKSDAQR | Major Facilitator Superfamily (MFS) | 27.45 |

| | | | | | | |
|----------|----------|----------|--------------|--|--|-------|
| 1.00E+04 | 4.00E+06 | 3.00E+06 | GABP01104797 | KQTFSDSDEEEKPPAK | RRM domain containing protein | 51.97 |
| 7.00E+04 | 3.00E+05 | 1.00E+05 | GABP01028262 | KRPLEDGEAEADSPPGGEAEDGEAGEDGK | RRM domain containing protein | 89.94 |
| 1.00E+04 | 3.00E+05 | 3.00E+05 | GABP01054190 | LAPHSAA S PEPSRPRSPEPAER | Unknown | 18.94 |
| 1.00E+04 | 4.00E+05 | 1.00E+05 | J0717160 | LATYKVPQELEAVPELPR | AMP-dependent synthetase and ligase | 18.68 |
| 1.00E+04 | 2.00E+05 | 4.00E+05 | J0761509 | LDQITLPV S MLR | serine/threonine protein kinase related protein | 35.9 |
| 3.00E+05 | 2.00E+06 | 3.00E+06 | GABP01051799 | LDS N EDVRPSAGMR | Unknown | 35.01 |
| 1.00E+04 | 2.00E+06 | 1.00E+06 | GABP01086664 | LENVMVDME S PKR | Protein kinase | 36.88 |
| 1.00E+04 | 9.00E+05 | 6.00E+05 | J0743538 | LEPEQSGS S PPAR | serine/threonine protein phosphatase | 20.79 |
| 1.00E+04 | 4.00E+05 | 3.00E+05 | GABP01025693 | LGASLALFDFAHMGASLSLR | Unknown | 16.76 |
| 1.00E+04 | 1.00E+05 | 8.00E+04 | GABP01046365 | LGEL S AMIEATK | Unknown | 17.05 |
| 1.00E+04 | 6.00E+05 | 3.00E+05 | J0725591 | LGPMYGERSDDEGS G AEVK | Unknown | 56.81 |
| 3.00E+05 | 8.00E+05 | 1.00E+06 | GABP01078304 | LGSLNPSITPALLEAS E ALPK | chalcone and stilbene synthase domain-containing protein | 33.91 |
| 1.00E+04 | 2.00E+05 | 2.00E+05 | J0714359 | LKF S PMAMQQPQAPK | hypothetical protein Pmar_PMAR024059 | 21.54 |
| 1.00E+04 | 2.00E+05 | 2.00E+05 | GABP01029502 | LLFAGPAV T RRTLAGE S LAPSTSSPASTPSS SAR | Cell adhesion function related protein | 44.16 |
| 1.00E+04 | 9.00E+05 | 3.00E+05 | GABP01059430 | LLGGSLAGLYRPGAG S K | Dihydrogluconate reductase domain containing | 15.5 |
| 1.00E+04 | 3.00E+05 | 3.00E+05 | GABP01058661 | LLLT T TTTACVCLSGLQHACK | Thrombospondin type 1 repeat domain containing | 20.65 |
| 9.00E+05 | 2.00E+06 | 2.00E+06 | J0720445 | LL S RDEESGPTAK | predicted protein | 23.59 |
| 1.00E+04 | 2.00E+06 | 2.00E+06 | GABP01012072 | LNALLPAV S SRPMAPSGSEVIRTFAAHAK | Unknown | 24.86 |
| 1.00E+04 | 5.00E+05 | 4.00E+05 | J0722680 | LPGGAMPAP T MTQHMS Y IQKMNPIEAM VTAVLVK | cAMP-dependent protein kinase regulatory subunit | 15.57 |
| 1.00E+04 | 8.00E+04 | 1.00E+05 | GABP01061399 | LPSGLPDCT T P T ASEELGAWR | Unknown | 55.81 |
| 1.00E+04 | 3.00E+05 | 3.00E+05 | GABP01018605 | LPS W TGLFR | alpha-glucan water dikinase | 42.73 |
| 1.00E+04 | 2.00E+05 | 5.00E+05 | GABP01076806 | LPT A LSR | Armadillio like domain containing | 20.96 |
| 1.00E+04 | 1.00E+05 | 4.00E+05 | J0731589 | LPTLVVALLAP S | ZOG-Fe(II) oxygenase superfamily protein | 15.96 |
| 1.00E+04 | 3.00E+05 | 3.00E+05 | GABP01069110 | LQVMIMIVWSASHDLL S ASIRAWK | Spliceosome complex protein | 20.29 |
| 1.00E+04 | 6.00E+05 | 3.00E+05 | J0718357 | LRS V CMHVLLLK | ADP-ribosylation factor | 18.66 |
| 1.00E+04 | 1.00E+05 | 9.00E+04 | J0727148 | LSEAPQKPAVQQ A SESATAAADEALAILR | predicted protein | 15.45 |
| 1.00E+04 | 5.00E+05 | 5.00E+05 | GABP01081887 | LSFKDPDAN S DDGN S DAEEVPKPR | XAP5 Nuclear protein | 87.56 |
| 1.00E+04 | 6.00E+05 | 2.00E+05 | J0754791 | L S FMLK | cytosolic tRNA-Ala synthetase | 16.51 |
| 2.00E+04 | 5.00E+05 | 2.00E+04 | J0705826 | LSLQPVFV S LAR | Unknown | 38.88 |
| 1.00E+04 | 6.00E+05 | 4.00E+05 | J0732494 | LTSSFTLTEAP S K | calcium-dependent protein kinase | 15.67 |
| 1.00E+04 | 3.00E+05 | 2.00E+05 | GABP01110361 | LVLGDSPADLEIREAGL S R | Dynein heavy chain family protein | 47.86 |
| 1.00E+04 | 5.00E+05 | 2.00E+05 | J0705182 | MEDEER S PTGPPPSK | NFX1-type zinc finger-containing protein 1 | 42.84 |
| 1.00E+04 | 2.00E+05 | 4.00E+05 | J0707767 | MIG T PCQRALSTAAVAQAQPR | tetratricopeptide repeat-containing protein | 22.82 |
| 2.00E+05 | 5.00E+05 | 4.00E+05 | GABP01008094 | MISDVDDDG S GTIGYEEFLK | Centrin | 41.61 |
| 1.00E+04 | 1.00E+05 | 4.00E+05 | J0703826 | M SRPV S WNFRD S LTAMTLSCAR | isocitrate dehydrogenase, NADP-dependent | 16.27 |
| 1.00E+05 | 4.00E+05 | 4.00E+05 | GABP01087383 | MTKP S LTAGPAVLR | Unconventional myosin | 34.88 |

| | | | | | | |
|----------|----------|----------|--------------|---|---|-------|
| 1.00E+04 | 8.00E+06 | 2.00E+06 | J0700473 | NATALLGN S GVQP | salivary gland secretion 1 | 28.55 |
| 7.00E+04 | 2.00E+05 | 3.00E+05 | GABP01106531 | NGSPSLLTSSTGLLCGSASSG S VR S CLR | Unknown | 24.31 |
| 1.00E+04 | 5.00E+05 | 2.00E+05 | GABP01084598 | NLT C KLVMVLSNGR | Unknown | 28.31 |
| 1.00E+04 | 2.00E+05 | 2.00E+05 | GABP01086127 | NSAGPLAVRPTHVQAAC T MR | zeaxanthin epoxidase, chloroplast precursor | 16.22 |
| 1.00E+04 | 1.00E+06 | 2.00E+06 | GABP01025477 | PAC T AAQAASQPPRRPTLLSFAGAMAAK | Unknown | 16.5 |
| 1.00E+04 | 5.00E+05 | 1.00E+05 | J0728952 | PQLSPTCLPF S AQAGPAQVAMPAK | eukaryotic translation initiation factor 5B, putative | 18.41 |
| 1.00E+04 | 2.00E+06 | 4.00E+06 | GABP01036474 | PSVIAAG S R | Unknown | 27.42 |
| 1.00E+04 | 8.00E+05 | 5.00E+05 | GABP01025605 | QCGTIT S WNMARAGK | CSP/OB fold domain containing protein | 29.36 |
| 7.00E+04 | 4.00E+05 | 1.00E+06 | GABP01050228 | QGEASALLT M SNLQAN S QAPTDALRSAR | TPR repeat containing protein | 15.95 |
| 1.00E+04 | 6.00E+06 | 2.00E+06 | GABP01114065 | QGLRPAP S GIGG S TRPCSS R SCGR | farnesyltransferase/geranylgeranyl transferase type 1 alpha subunit | 15.01 |
| 1.00E+04 | 1.00E+05 | 6.00E+04 | GABP01097283 | QLLAGLMVWPGGTWRS A SAGTPG T APTK | Unknown | 23.05 |
| 1.00E+04 | 1.00E+06 | 7.00E+05 | GABP01007088 | QL V ELEMEQK | Rna polymerase associated protein Leo1 | 17.5 |
| 1.00E+04 | 1.00E+05 | 3.00E+05 | GABP01018625 | QRPGPSAPPAPASGRPSVVGYPGMPPK S PANR | hypothetical protein HCAG_09240 | 37.72 |
| 1.00E+04 | 1.00E+05 | 8.00E+04 | GABP01028980 | Q SPQPPIPTTSHQRPPCQTRSALLSSAR | Solute carrier member 35 e4 | 18.65 |
| 1.00E+04 | 3.00E+05 | 3.00E+05 | J0721849 | QV S VPYYSSDLTR | glutamate decarboxylase | 24.01 |
| 1.00E+04 | 5.00E+05 | 2.00E+05 | GABP01034927 | RAGDVLES I AHVGADFQELWQGAAAHPRPAAGR | 2-hydroxyacylsphingosine 1-beta-galactosyltransferase-like | 26.81 |
| 1.00E+04 | 1.00E+06 | 6.00E+05 | GABP01060639 | RAIHGVPVAPQ S IPLP | Unknown | 15.91 |
| 4.00E+05 | 1.00E+06 | 4.00E+05 | GABP01096447 | REGAED S EEEEEPK | Unknown | 86.13 |
| 1.00E+04 | 2.00E+05 | 2.00E+05 | GABP01038926 | RGADGGDG A SG S ESEGEVER | Formin binding | 48.8 |
| 1.00E+04 | 3.00E+06 | 3.00E+06 | GABP01015275 | RGDD S GGEEEQFR | DNAJ heat shock N-terminal domain-containing protein | 47.65 |
| 1.00E+04 | 3.00E+05 | 4.00E+05 | GABP01007184 | RGGPQCQ W ASLWALR | PREDICTED: similar to protein phosphatase 1G variant isoform 5 | 15.57 |
| 1.00E+04 | 4.00E+05 | 3.00E+05 | GABP01017850 | RICLWIL S MFVLI | alpha-tubulin, partial | 16.07 |
| 1.00E+04 | 4.00E+05 | 1.00E+05 | GABP01055853 | RITLWPR | SRC homology 3 domain containing | 36.76 |
| 1.00E+04 | 2.00E+06 | 1.00E+06 | J0742036 | RNS S IATPVSNDR | cAMP-dependent protein kinase regulatory subunit | 56.03 |
| 1.00E+05 | 6.00E+05 | 1.00E+05 | J0742036 | RNS S IATPVSNDR | cAMP-dependent protein kinase regulatory subunit | 41 |
| 3.00E+05 | 8.00E+05 | 1.00E+06 | GABP01001006 | RPD S PGLQDR | Unknown | 37.44 |
| 4.00E+04 | 2.00E+07 | 1.00E+07 | J0754267 | RPGPHAVGRPR T QGGR | Protein kinase domain containing protein | 21.31 |
| 1.00E+04 | 2.00E+06 | 2.00E+06 | J0765479 | RPLLVTAAATLLNFC S LCLMP P TTK | Unknown | 28.44 |
| 7.00E+04 | 4.00E+05 | 2.00E+05 | GABP01047795 | RPP S PEGGGAEGPDDAG S PHAGNR | Unknown | 80.19 |
| 1.00E+04 | 6.00E+05 | 6.00E+05 | GABP01045657 | RPSGQAP S PAR | Unknown | 44.29 |
| 1.00E+04 | 9.00E+05 | 7.00E+05 | GABP01034357 | RQISSSP S IAAPAVR | Zn finger domain containing protein | 24.29 |
| 1.00E+04 | 2.00E+06 | 6.00E+05 | J0701179 | RSDVPLG A SAGASTLMSPCGPCTEK | Unknown | 15.5 |
| 6.00E+04 | 6.00E+05 | 1.00E+06 | GABP01035906 | R SPLPPGPPLPPGSR | Unknown | 35.45 |
| 1.00E+04 | 7.00E+05 | 4.00E+05 | GABP01029422 | R SSRPRLAPVRPRPAGADGRP | RNA binding protein NOVA-2 | 16.6 |
| 1.00E+04 | 1.00E+06 | 1.00E+06 | J0693972 | RT A SNPAGMLLLMPSLK | hypothetical protein TGME49_027840 | 19.51 |
| 9.00E+04 | 2.00E+05 | 3.00E+05 | GABP01036474 | SAAPVETPADGAVTAA A SPR | Unknown | 57.73 |
| 1.00E+04 | 3.00E+05 | 3.00E+05 | GABP01100888 | SATQAPLVSTDPIDEATEFAGPP S SAVK | ARF GTPase activating protein | 65.55 |

| | | | | | | |
|----------|----------|----------|--------------|--|---|-------|
| 1.00E+04 | 3.00E+05 | 2.00E+05 | GABP01027848 | SCTVGGCPHTLPPPPRPR | SET domain protein | 32.11 |
| 1.00E+04 | 3.00E+05 | 3.00E+05 | GABP01063423 | SDSPDALGDFNSGQGGGR | Unknown | 88.71 |
| 1.00E+04 | 5.00E+06 | 3.00E+06 | GABP01112644 | SFATLQPSTYGD SKL | Unknown | 17.05 |
| 1.00E+04 | 3.00E+05 | 2.00E+05 | GABP01107575 | SFGSGTNLADLASQSR | delta-12 oleate desaturase | 22.17 |
| 1.00E+04 | 9.00E+05 | 4.00E+05 | GABP01038668 | SGPQTVVPCGSKTMT PQAPR | Hect E3 Ubiquitin ligase | 15.28 |
| 1.00E+04 | 6.00E+05 | 3.00E+05 | GABP01082788 | SGTLGLQGPDSPGASAGR | Unknown | 49.77 |
| 1.00E+05 | 2.00E+06 | 1.00E+06 | GABP01082788 | SGTLGLQGPDSPGASAGR | Unknown | 41.1 |
| 1.00E+04 | 2.00E+05 | 2.00E+05 | GABP01061823 | SLACAIPEVAREASSVAGTR | Peptidase M16 domain containing protein | 35.25 |
| 2.00E+06 | 9.00E+06 | 5.00E+06 | GABP01049965 | SLATSNPGESPQGPKEGK | Pumilio | 67.07 |
| 1.00E+04 | 5.00E+06 | 6.00E+06 | GABP01025983 | SLCCIQSSNDPWNLPK | Unknown | 25.83 |
| 1.00E+04 | 5.00E+05 | 2.00E+05 | GABP01025983 | SLCCIQSSNDPWNLPK | Unknown | 38.26 |
| 1.00E+04 | 3.00E+05 | 5.00E+05 | GABP01014465 | SLRPTASRQCAAR | step II splicing factor SLU7, putative | 17.49 |
| 1.00E+04 | 2.00E+05 | 4.00E+05 | GABP01004920 | SLSFDSPSGPPSK | Unknown | 35.99 |
| 1.00E+04 | 2.00E+05 | 8.00E+04 | JO719720 | SLVEEGAPSVFVLADK | Unknown | 15.23 |
| 1.00E+04 | 2.00E+06 | 2.00E+06 | GABP01091939 | SLYAQLSSAVCSLVSRGTGVSHAMPTKPR | Unknown | 18.41 |
| 1.00E+04 | 3.00E+05 | 4.00E+05 | GABP01113699 | SNACPPPALMAASVSSPSQALTALQVTML GVMWR | Unknown | 17.12 |
| 4.00E+04 | 1.00E+05 | 2.00E+05 | GABP01113699 | SNACPPPALMAASVSSPSQALTALQVTML GVMWR | Unknown | 29.15 |
| 1.00E+04 | 1.00E+05 | 1.00E+05 | GABP01033307 | SPSRDGSPPNDWIVNMLIDR | Unknown | 16.81 |
| 1.00E+04 | 7.00E+04 | 2.00E+05 | JO710259 | SRPAVASRSQLQR | Unknown | 22.72 |
| 1.00E+04 | 4.00E+05 | 8.00E+05 | JO706757 | SRTSTSRPPR | Unknown | 25.16 |
| 4.00E+04 | 1.00E+05 | 1.00E+05 | GABP01011206 | SSGRTTASSTPR | mitochondrial ATP synthase F1 alpha subunit-like protein 1 | 20.26 |
| 1.00E+04 | 7.00E+05 | 1.00E+06 | GABP01034207 | SSVQQAFTTR | serine/threonine protein kinase related | 23.32 |
| 1.00E+04 | 1.00E+06 | 4.00E+05 | JO695593 | STAMSPKEIEGR | hypothetical protein CHLNCRAFT_137920 | 24.26 |
| 1.00E+04 | 3.00E+05 | 2.00E+05 | JO721667 | STGRPAQPAPASATALKMSAPR | Unknown | 16.81 |
| 1.00E+04 | 3.00E+06 | 4.00E+05 | GABP01012985 | STSNPYDERPVGK | calcium-binding protein | 26.16 |
| 1.00E+04 | 2.00E+05 | 1.00E+05 | GABP01092058 | STSRPPAQSSGGTSPVAER | Unknown | 29.68 |
| 1.00E+04 | 5.00E+05 | 8.00E+05 | JO725087 | STWEPQGIAS PAGWK | apicomplexan-conserved protein | 51.44 |
| 1.00E+04 | 4.00E+06 | 1.00E+06 | JO704770 | TAALASTVPESARMNSWYCSTLPHSVSR | CG7139, isoform A | 17.26 |
| 1.00E+04 | 6.00E+05 | 5.00E+05 | GABP01107690 | TALTTVAPPAALLQVR | Unknown | 17.48 |
| 1.00E+04 | 3.00E+05 | 2.00E+05 | GABP01028183 | TDSPDAESESDEEPPKK | Structural maintenance of Chromosomes SMC4 | 33.79 |
| 1.00E+04 | 3.00E+05 | 1.00E+05 | GABP01008016 | TGSVCEDSTITLPLGLGK | ML superfamily protein | 31.3 |
| 1.00E+04 | 2.00E+06 | 1.00E+05 | GABP01098060 | TNGHVLPVSQYLQHWRR | Unknown | 15.74 |
| 1.00E+04 | 2.00E+05 | 7.00E+05 | GABP01111212 | TPAAGARPRSLR | Unknown | 22.46 |
| 2.00E+05 | 8.00E+05 | 6.00E+05 | GABP01059771 | TPASLAGSPVPVLGASR | Unknown | 26.62 |
| 1.00E+04 | 3.00E+05 | 3.00E+05 | GABP01019516 | TPSTTATEESDGEEDDQLTHLTAPAEPA QPK | hypothetical protein CHLREDRAFT_99865 | 49.94 |
| 1.00E+04 | 6.00E+06 | 4.00E+06 | GABP01048268 | TQSCRWPWPCGAPSTCARWTCR | Unknown | 18.62 |
| 1.00E+04 | 6.00E+05 | 1.00E+06 | GABP01025130 | TSLPAGSASSPLTSLR | hypothetical protein | 32.38 |

| | | | | | | |
|----------|----------|----------|--------------|---|---|--------|
| 1.00E+04 | 5.00E+05 | 8.00E+05 | GABP01092376 | TSNGTTSVALRTPK | Cell division cycle protein | 20.88 |
| 1.00E+04 | 3.00E+05 | 4.00E+05 | GABP01092376 | TSNGTTSVALRTPK | Cell division cycle protein | 39.37 |
| 1.00E+04 | 3.00E+07 | 5.00E+05 | GABP01084233 | TSPCCTVARSTACTPAEAVR | Unknown | 29.55 |
| 1.00E+04 | 3.00E+06 | 2.00E+06 | GABP01024322 | TTARLVCAGPLALLPSASSPVAR | DDE superfamily endonuclease containing protein | 15.37 |
| 1.00E+04 | 4.00E+05 | 4.00E+05 | GABP01058507 | TTSEIVQAPKPSLPGAGSR | Unknown | 39.74 |
| 1.00E+04 | 1.00E+05 | 4.00E+05 | GABP01086031 | TTTVSAQASTKAAASMALR | SET domain containing protein | 27.93 |
| 1.00E+04 | 2.00E+05 | 5.00E+05 | GABP01049570 | TWTLCGTPEYLAPEIIQSK | cAMPdependent protein kinase | 83.25 |
| 1.00E+04 | 7.00E+05 | 1.00E+06 | J0756760 | TYTLCGTPEYIAPEVLLNK | protein kinase | 44.37 |
| 1.00E+04 | 3.00E+06 | 5.00E+06 | GABP01036663 | TYTLCGTPEYIAPEVLLNK | cAMP dependent protein kinase Catalytic | 97.02 |
| 1.00E+04 | 3.00E+05 | 5.00E+05 | J0746453 | VAEPTGTDTPR | Unknown | 44.44 |
| 1.00E+05 | 5.00E+05 | 4.00E+05 | GABP01040479 | VASSMMDPESPRSPGK | Unknown | 34.29 |
| 1.00E+04 | 5.00E+05 | 7.00E+05 | GABP01093776 | VFYHLEEDDAR | proteasome 26S non-ATPase subunit 1 | 33.88 |
| 1.00E+06 | 5.00E+06 | 3.00E+06 | J0733631 | VGDEQGPSVLDAQDPNYDSEETAKK | conserved hypothetical protein | 79.7 |
| 1.00E+04 | 5.00E+05 | 6.00E+05 | J0719542 | VGNSDLGEMAGGNSPDR | Splicing factor, arginine/serine-rich, putative | 64.46 |
| 1.00E+04 | 7.00E+05 | 3.00E+05 | GABP01020442 | VLVTGAGGRTGSMVLQK | isomerase | 23.61 |
| 1.00E+04 | 2.00E+05 | 2.00E+05 | J0719542 | VPAHGGAPQVDKVEEMAFSEED | Immediate-early protein | 39.98 |
| 1.00E+05 | 3.00E+05 | 3.00E+05 | GABP01018625 | VQTLMLQPVLMTSPGSPGK | hypothetical protein HCAG_09240 | 64.4 |
| 1.00E+04 | 4.00E+05 | 3.00E+05 | GABP01101024 | VRSAITLTPR | Serine protease family protein | 21.46 |
| 1.00E+04 | 2.00E+05 | 2.00E+05 | GABP01094006 | VSAQHDAAGSGDDAGEAR | Unknown | 92.63 |
| 1.00E+04 | 5.00E+05 | 3.00E+05 | GABP01053847 | VTDDSPAAPPQK | Amino acid transporter | 26.47 |
| 1.00E+04 | 1.00E+05 | 2.00E+05 | GABP01093834 | VTVAEAPSLWPWASLAMPASAAATEARK | Unknown | 27.27 |
| 1.00E+04 | 2.00E+05 | 2.00E+05 | GABP01025130 | WADATPTMAFTPAEDSR | hypothetical protein | 18.6 |
| 1.00E+04 | 2.00E+05 | 4.00E+05 | GABP01041205 | WADCSEDEEDER | Unknown | 45.58 |
| 1.00E+04 | 8.00E+05 | 1.00E+06 | GABP01094643 | WDEVESDPDEPCPDVAK | Unknown | 92.5 |
| 3.00E+05 | 1.00E+06 | 1.00E+06 | J0706453 | WGAVNPGSPGGAGGWR | fibrinogen A-alpha chain | 47.91 |
| 1.00E+04 | 4.00E+05 | 2.00E+05 | GABP01020598 | WGPSARAAGRPSPGR | PAP/25A associated domain-containing protein | 23.24 |
| 1.00E+04 | 4.00E+05 | 5.00E+05 | J0743481 | WPGLLVLAQSLTPATSPAR | saccharopine dehydrogenase | 32.08 |
| 1.00E+04 | 3.00E+05 | 4.00E+04 | GABP01068279 | WTKATPLAQGGDEACSQLPASTWGK | Selenoprotein | 27.98 |
| 1.00E+04 | 2.00E+05 | 3.00E+05 | GABP01000248 | YAPINVLDSDDEADAPLPPPPK | Unknown | 104.21 |
| 2.00E+05 | 5.00E+05 | 6.00E+05 | J0706783 | YGEDSGDEILR | unnamed protein product | 37.52 |
| 3.00E+05 | 3.00E+06 | 1.00E+06 | GABP01010423 | YPTPSNRQSTVQPASQR | Unknown | 22.62 |
| 1.00E+04 | 7.00E+04 | 7.00E+04 | GABP01043687 | YPVSLSTLNTSGNDVPNSSCGNYWSAVPR CAQEGLVPSLLAGSSK | Unknown | 17.29 |

Table 5.ST3. Differential RNA expression in cysts

| Sequence Number | Accession | Mean Read Count | Log 2 difference | P value | Sequence ID | |
|---------------------------|-----------|-----------------|------------------|----------|--|------|
| Plastid | | | | | | |
| JO730904 | | 160 | -3.9 | 2.00E-04 | ATP synthase subunit alpha | atpA |
| JO692713 | | 214693 | -3.5 | 4.00E-06 | ATP synthase CF1 alpha | atpA |
| JO692692 | | 257356 | -4.2 | 1.00E-07 | ATP synthase CF1 beta | atpB |
| JO692682 | | 82772 | -3.9 | 4.00E-07 | ATP synthase beta | atpB |
| JO692702 | | 54054 | -4.8 | 1.00E-09 | PSI p700 chlorophyll a apoprotein a1 | psaA |
| JO767270 | | 146801 | -4.5 | 8.00E-09 | PSI p700 chlorophyll a apoprotein a1 | psaA |
| JO692700 | | 60170 | -4.2 | 6.00E-08 | PSI p700 chlorophyll a apoprotein a1 | psaA |
| JO692626 | | 26393 | -3.8 | 1.00E-06 | PSI chain b | psaB |
| JO692701 | | 62540 | -3.7 | 2.00E-06 | PSI chain b | psaB |
| JO692681 | | 83938 | -3.6 | 3.00E-06 | PSI chain b | psaB |
| JO692622 | | 26213 | -3.3 | 2.00E-05 | PSI chain b | psaB |
| JO692697 | | 61271 | -4.1 | 1.00E-07 | PSI chain b | psaB |
| JO692693 | | 70845 | -3.5 | 5.00E-06 | PSI p700 apoprotein a2 | psaB |
| JO692705 | | 45482 | -4.2 | 6.00E-08 | PSII CP47 apoprotein | psbB |
| JO692686 | | 74511 | -4.1 | 2.00E-07 | PSII CP47 apoprotein | psbB |
| JO692629 | | 26973 | -3.9 | 4.00E-07 | PSII CP47 apoprotein | psbB |
| JO692706 | | 45146 | -3.9 | 6.00E-07 | PSII CP47 apoprotein | psbB |
| JO692710 | | 233663 | -3.9 | 6.00E-07 | PSII CP43 chlorophyll apoprotein | psbC |
| JO692645 | | 132238 | -3.4 | 7.00E-06 | PSII CP43 chlorophyll apoprotein | psbC |
| JO692709 | | 45233 | -3.2 | 3.00E-05 | Cytochrome b6 | petB |
| JO692630 | | 28544 | -3.5 | 5.00E-06 | Cytochrome b6f complex subunit IV | petD |
| Cell Wall/Membrane | | | | | | |
| JO696533 | | 1056 | -3.9 | 2.00E-06 | Cell surface protein p43 | |
| JO716857 | | 16619 | -3.2 | 3.00E-05 | Cell surface protein p43 | |
| JO754823 | | 12999 | -3.2 | 3.00E-05 | Cell surface protein p43 | |
| JO729217 | | 15149 | -2.6 | 5.00E-04 | Cell surface protein p43 | |
| JO751668 | | 1083 | 3.1 | 4.00E-05 | Chitin binding protein | |
| JO741654 | | 1244 | 2.7 | 3.00E-04 | Hemagglutinin hemolysin-related protein | |
| JO759225 | | 2007 | -6 | 2.00E-12 | Cell wall associated hydrolase | |
| JO724223 | | 528 | -5.4 | 3.00E-09 | Cell wall associated hydrolase | |
| JO741027 | | 117 | -5.1 | 2.00E-05 | Cell wall associated hydrolase | |
| JO711384 | | 1309 | -3.1 | 1.00E-04 | Cell wall associated hydrolase | |
| JO710531 | | 8903 | -6.3 | 3.00E-14 | Cell wall associated hydrolase | |
| Miscellaneous | | | | | | |
| JO732017 | | 1756 | -4 | 9.00E-07 | pg1 protein | |
| JO699453 | | 3010 | -7.7 | 9.00E-18 | pg1 protein | |
| JO702044 | | 1848 | -4.8 | 5.00E-09 | pg1 protein | |
| JO749682 | | 155 | -4.6 | 2.00E-05 | Glycosyl transferase | |
| JO694867 | | 3677 | -4.5 | 1. E-08 | Translation elongation factor like protein | |
| JO747847 | | 145 | -9.1 | 7.00E-10 | Dual specificity phosphatase | |
| JO703232 | | 852 | -3.8 | 5.00E-06 | rRNA intron encoded homing | |

| | | | | | |
|----------|-----|------|----------|-------------------------------------|--|
| | | | | endonuclease | |
| JO766746 | 76 | -6.2 | 2.00E-05 | Helicase like transcription factor | |
| JO705601 | 162 | -7.7 | 2.00E-09 | Uracil DNA glycosylase | |
| JO693202 | 56 | -6.7 | 9.00E-05 | is5 transposase and trans-activator | |

5.6. Acknowledgements

We thank L. Pelletier for assistance with the electron microscope, and E. Bonneil at the Institut de Recherche en Immunologie et Cancerologie for the mass spectrometry sequencing. This work was supported by the National Science and Engineering Research Council of Canada (NSERC).

CHAPTER 6 – GENERAL DISCUSSION

6.1. General Discussion

The dinoflagellates are a poorly characterized group of microbial eukaryotes, somewhat surprising as they display a number of fascinating biological features including bioluminescence, formation of red tides, and symbiosis with corals. The ecology of dinoflagellates has been the primary emphasis of research. However, several aspects of their cell biology, particularly the strange chromatin organization, an apparent preference for translational regulation of gene expression and a widespread role of the circadian clock in regulation of cell behaviour has stimulated a renewed interest in this system. As the physiological processes, including blooming under nutrient rich conditions and toxin production, all presumably result from regulation of gene expression, to understand these processes it is necessary to gain more insight into the basic molecular biology of dinoflagellates. Proper knowledge of molecular mechanisms that regulate the dynamics of gene expression will not only help to understand the changes in physiology but may also provide information on strategies for regulating gene expression that differ from typical model eukaryotes. *Lingulodinium polyedrum* has been the centre of interest for its spectacular nightly bioluminescence and has been extensively studied to understand the biochemistry behind the dinoflagellate circadian rhythms [134, 587].

L. polyedrum has a typical dinoflagellate nucleus (a “dinokaryon”) that contains 200 pg of DNA folded into liquid crystalline and permanently condensed chromosomes. It is the dinoflagellate in which circadian clock control over gene expression is most widely studied. It is unfortunate that *L. polyedrum* as other dinoflagellate have to date not been amenable to transformation or mutational analysis, but the advancement of technologies like RNA-seq and mass spectrometry provided some alternate approaches. In particular, deep sequencing of the transcriptome and high throughput proteome approaches have been particularly useful in the studies that I performed in *L. polyedrum*.

Any global gene expression studies require a reliable database for comparison and analysis. To produce this database, the RNA-seq technology from Illumina was used to

sequence 76 bp from each side of the 300bp cDNA fragments generated from *L. polyedrum* RNA. These paired end reads were then assembled by a *de novo* based technique into 74,655 transcripts of minimum length 300 bp, the first comprehensive transcriptome in *L. polyedrum*. Sequence homology searches were then used to determine the extent of presence of a number of different biochemical pathways as a means of assessing the completeness of the transcriptome [165].

My own interest in sequence homology searches was to unveil atleast some of the transcripts involved in basic gene expression and related pathways. I found it remarkable that DNA binding domains were scarce in the transcriptome, and of those that were present, the majority contained the cold-shock domain (CSD). This is a peculiar situation, as CSD are not as common in other eukaryotes. It is also interesting to note that while classified as transcriptional regulators, many studies in other eukaryotes as well as bacteria assign a general translational regulatory role to CSD proteins. In addition, dinoflagellates do not have a TATA-binding protein TBP, unlike all other eukaryotes tested so, but have instead a TBP-like protein called TLF. First identified over ten years ago, the actual role of TLF in transcription remains obscure [179]. In contrast to the transcriptional regulators, the *L. polyedrum* transcriptome contains a full suite of translation factors, which are, in general, quite well conserved. One noteworthy exception is the absence of eIF2B, the guanine exchange factor required to charge eIF2 with GTP. In other eukaryotes, the phosphorylation of eIF2 alpha at serine-151 results in higher affinity of eIF2 binding to eIF2B. This has as a result the sequestering of eIF2B, and a reduction of global translation rates, a mechanism with critical implications in mammals under stress [588]. As yet, it seems this mechanism might not have a big impact in regulating global translation rates in dinoflagellates.

Gene arrangement can also be very important for regulation of transcription. Dinoflagellates contain a unique arrangement often involving tandem repeats of multiple copies of the same gene. If each of these copies is transcribed separately, each repeat unit would contain its own promoter element. However, when the intergenic regions of several tandem repeat genes was analysed, no significant and consistent sequence conservation was found. This, in combination with the discovery that dinoflagellates perform SL *trans*

splicing, led to the hypothesis of polycistronic transcript formation and subsequent processing. The biological precedent for this proposal is in the *Trypanosoma*, where SL *trans* splicing carves individual transcripts out of a long polycistronic transcript. In theory, this should result in equivalent amounts of coding sequence and intergenic region sequence in the transcriptome. However, aligning the RNA-seq reads to the genomic copies of PCP and LCF showed 5 to 36 thousand more reads for the coding compared to the intergenic regions for these two genes. This, in combination with other evidence (presented in chapter 3), the presented evidences argue against the formation of polycistronic transcripts in dinoflagellates. It is also interesting to note that mapping individual reads back to the coding sequences can also be used to assess sequence diversity at each nucleotide position. Here, high copy number genes are remarkably well conserved, suggestive of a mechanism similar to that preserving ribosomal RNA sequences. However, in contrast to these high copy number tandem array genes, when genes potentially acquired through horizontal gene transfer (HGT) are examined, a large number of mutations with respect to the original sequence are expected. These mutations occur in order to adjust the GC content of the original sequence to the GC preference of the new host cell, and these changes may have been allowed because at the time of the HGT the sequence was presumably present in only a single copy.

Epigenetic regulation of gene expression relies, atleast in part, on the changes in chromatin structure brought about by modification of histone proteins [397]. However, neither protein extraction protocols nor DNA spreads under the electron microscope showed the presence of histones or nucleosomes in dinoflagellates. Despite this, the *L. polyedrum* transcriptome has several different copies of each of the four core histones as well as a set of enzymes and chaperone proteins associated with histone modification and nucleosome assembly, respectively. All these sequences are predicted to form functional proteins, providing strong presumptive evidence that they may be selected for. However, despite the fact that neither immunological methods nor LC-MS/MS of acid extracted protein fractions showed any histone proteins, the presence of trace amounts of histones in particular genomic regions is still an intriguing possibility.

Even after a protein is synthesized, post translational modification such as phosphorylation /dephosphorylation can alter function and thus should be considered as important for regulation of gene expression. Protein phosphorylation also plays a crucial role in regulating the circadian system of *L. polyedrum*, as shown by studies using broad-specificity kinase inhibitors. I used the transcriptome sequence database to characterize the kinase families present in a dinoflagellate, which was not available before. Though not exhaustive but this analysis presented two interesting features. First, there is an overabundance of calcium dependant protein kinases (CDPKs) and second, all kinases involved in the clock in other organisms (CK1, CK2, AMPK and Gsk3 β) are present in the transcriptome. I was also interested in applying phosphoproteomic approaches to the dinoflagellates, the experiments that were previously impossible due to the lack of a comprehensive sequence database. The use of MS techniques to analyse a phosphoprotein enriched protein fraction allowed me to identify the phosphosites on over 500 phosphopeptides and compared their relative levels between different samples. Using the amino acid context to assign the phosphosites as target of particular kinase classes, I identified several different RNA binding proteins that were potential CK2 targets and differentially phosphorylated at the two times investigated. This is potentially interesting, as these differentially phosphorylated RBPs are good candidates for mediating the previously reported circadian control of proteins synthesis at these two time points [347]. This method could become a useful tool to explore the circadian regulation of translation in *L. polyedrum*.

The role of CSDs in regulating gene expression under cold temperature shock is a well-known phenomenon in eukaryotes as well as prokaryotes [351, 518]. I thus attempted to find out if the overabundance of CSD domains in *Lingulodinium* played any role in a cold shock response. I examined the transcriptomics, the proteomics, and in particular the phosphoproteomics of the cold-treated *Lingulodinium*. The first observation to note is that cold treatment induces the formation of cysts, metabolically dormant cells whose role is to resist adverse environmental conditions. Comparing transcriptome wide RNA levels between cysts and motile cells showed the plastid encoded RNAs to be significantly lower in the cysts, which at this time can be interpreted as a result of either preferential degradation, inhibition of transcription or a combination of both. However, it is worth mentioning that as

yet, we are unaware of the half life of these RNAs. I also noticed remarkable changes in the number and shape of chloroplasts in cysts using fluorescent microscopy. Surprisingly 2D PAGE analysis revealed no significant differences in the amounts of any of the cyst proteins, suggesting that CSD proteins may be involved in cellular responses rather than cold stress response. A major difference was noticed in the phosphoproteome, which showed significant hypophosphorylation in cysts. Furthermore, using the kinase prediction method developed previously, CK2 substrates appear to be the most hypophosphorylated. This again speaks to a potential role of CK2 in the dinoflagellate circadian system. I hypothesize that it is the profound changes in the phosphoproteome that are able to regulate the cellular physiology of the cells on a time frame consistent with that of encystment/excystment.

Taken together, the poor conservation of the transcription machinery and extremely low amounts of DNA binding domain all indicates that regulation of gene expression at transcription seems unlikely in dinoflagellates. In particular, epigenetic regulation of gene expression using histones seems not to play a major role in dinoflagellates, although the presence of trace amounts of histones is still possible that can regulate particular genes. On the other hand, there is strong evidence from my present work that phosphorylation/dephosphorylation of proteins will play a major role in determining the amount of active proteins in *Lingulodinium*. Evidence thus is consistent with the idea that *Lingulodinium* is rather unique eukaryotic system where translational and post-translational regulation of gene expression is predominant.

6.2. Future perspectives

The *L. polyedrum* transcriptome is one of the few comprehensive databases available for dinoflagellates, which can now be effectively used in large scale -omics studies. Apart from this, the sequences in the transcriptome will be useful to select and isolate by PCR those encoding proteins for which further biochemical experimentation will be required. CSD proteins will be the first to be investigated in order to explore the reason behind the expansion of this family in dinoflagellates. Furthermore, I have identified several other groups of proteins known to regulate translation in other eukaryotes. It will be important to

see to what extent they affect *Lingulodinium* gene expression. Lastly, I have provided here substantial indications that phosphoprotein profiling in dinoflagellates will be interesting in understanding gene regulation. In particular, the CK2 has been predicted to phosphorylate different RNA binding proteins at different times of the day, and thus further biochemical analysis will be required to confirm this potential role. Drug inhibition assays can be interesting option to find the effect of CK2 on the bioluminescence rhythm. However, a caveat to these experiments is that all the drugs are tested on mammalian systems and therefore their specificity requires reconfirmation in dinoflagellate system. If *Lingulodinium* CK2 influences its circadian clock as is observed in other eukaryotic systems, there is an opportunity to identify clock elements by thorough analysis of CK2 substrates on 24 hour scale phosphoproteome enrichment experiments. Some of the interesting targets already uncovered in this research can be used to test further using antibodies on Western blots.

7. Bibliographie

1. Fensome RA, MacRae, R.A. and Williams, G.L.: **Dinoflagellate Evolution and Diversity Through Time**. In: *Science Review 1994 & '95*. Edited by Fiander A. Dartmouth, Nova Scotia, Canada; 1996: 45 -50.
2. Cavalier-Smith T: **Cell diversification in heterotrophic flagellates**. . Oxford, UK.: Clarendon Press; 1991.
3. Gajadhar AA, Marquardt WC, Hall R, Gunderson J, Ariztia-Carmona EV, Sogin ML: **Ribosomal RNA sequences of Sarcocystis muris, Theileria annulata and Cryptosporidium parvum reveal evolutionary relationships among apicomplexans, dinoflagellates, and ciliates**. *Mol Biochem Parasitol* 1991, **45**(1):147-154.
4. Fast NM, Xue L, Bingham S, Keeling PJ: **Re-examining alveolate evolution using multiple protein molecular phylogenies**. *J Eukaryot Microbiol* 2002, **49**(1):30-37.
5. Moestrup OaD, N.: **On dinoflagellate phylogeny and classification**. New York: CRCPress; 2007.
6. Steidinger KAaT, K. : **Dinoflagellates**. New York: Academic Press; 1996.
7. Zingmark RG: **Sexual reproduction in the dinoflagellate Noctiluca miliaris Suriray**. *J Phycol* 1970, **6**:122-126.
8. Lucas IAN: **Observations on Noctiluca scintillans Macartney (Ehrenb.) (Dinophyceae) with notes on an intracellular bacterium**. *J Plank Res* 1982, **4**:401-409.
9. Eddy S: **The Fresh-water Armored or Thecate Dinoflagellates**. *Transactions of the American Microscopical Society* 1930, **XLIX**(4):277-321.
10. De Salas MF, Laza-Martínez, A. and Hallegraeff, G.M.: **ONovel unarmored dinoflagellates from the toxigenic family Kareniaceae (Gymnodiniales): Five new species of Karlodinium and one new Takayama from the Australian sector of the southern ocean**. *J Phycol* 2007, **44**(1):241 - 257.
11. JEFFREY SWaH, F.T.: **Photosynthetic pigments of symbiotic dinoflagellates (Zooxanthellae) from corals and clams**. *Biol Bull* 1968, **135**(1):149 - 165.
12. Gaines G, and Elbrachter, M. : **Heterotrophic nutrition**. Oxford, UK: Blackwell; 1987.
13. Goodson MS, Whitehead, L.F. and Douglas, A.E.: **Symbiotic dinoflagellates in marine Cnidaria: diversity and function**. *Hydrobiologia* 2001, **461**(1-3):79-82.
14. Coats DW: **Parasitic Life Styles of Marine Dinoflagellates**. *Journal of Eukaryotic Microbiology* 1999, **46**(4):402-409.
15. Stoecker DK: **Mixotrophy among Dinoflagellates**. *J Eukaryot Microbiol* 1999, **46**(4):397 -401.
16. Dodge JD: **in The Chromophyte Algae: Problems and Perspective**, vol. 38. Clarendon, Oxford; 1989.
17. Jeffrey SW: **in The Chromophyte Algae: Problems and Perspective**, vol. 38. Clarendon, Oxford; 1989.

18. Bachvaroff TR, Concepcion GT, Rogers CR, Herman EM, Delwiche CF: **Dinoflagellate expressed sequence tag data indicate massive transfer of chloroplast genes to the nuclear genome.** *Protist* 2004, **155**(1):65-78.
19. Hansen G, Moestrup, O. and Roberts, K.R.: **Light and electron microscopical observations on the type species of *Gymnodinium*, *G-fuscum* (Dinophyceae).** *Phycologia* 2000, **39**(5):365-376.
20. Daugbjerg N, Hansen, G., Larsen, J. and Moestrup, O. : **Phylogeny of some of the major genera of dinoflagellates based on ultrastructure and partial LSU rDNA sequence data, including the erection of three new genera of unarmoured dinoflagellates.** *Phycologia* 2000, **39**(4):302-317.
21. Fenchel T: **How dinoflagellates swim.** *Protist* 2001, **152**(4):329-338.
22. Miyasaka I, Nanba K, Furuya K, Nimura Y, Azuma A: **Functional roles of the transverse and longitudinal flagella in the swimming motility of *Prorocentrum minimum* (Dinophyceae).** *J Exp Biol* 2004, **207**(Pt 17):3055-3066.
23. Taylor DL: **Symbiotic marine algae: taxonomy and biological fitness.** Columbia University of South Carolina Press,; 1974.
24. McLaughlin JJA, and Zahl, P. A. : **Endozoic algae**, vol. 1. New York: Academic Press 1966.
25. Trench RK: **Dinoflagellates in non-parasitic symbioses.** Oxford: Blackwell Scientific,; 1987.
26. Stanley GD, and Swart, P.K.: **Evolution of the coral-zooxanthella symbiosis during the Triassic: a geochemical approach.** *Paleobiology* 1995, **21**:179-199.
27. Marshall AT: **Response: calcification rates in corals.** *Science* 1996, **274**(5284):117c-118c.
28. Scholin CA, Gulland F, Doucette GJ, Benson S, Busman M, Chavez FP, Cordaro J, DeLong R, De Vogelaere A, Harvey J *et al*: **Mortality of sea lions along the central California coast linked to a toxic diatom bloom.** *Nature* 2000, **403**(6765):80-84.
29. Flewelling LJ, Naar JP, Abbott JP, Baden DG, Barros NB, Bossart GD, Bottein MY, Hammond DG, Haubold EM, Heil CA *et al*: **Brevetoxicosis: red tides and marine mammal mortalities.** *Nature* 2005, **435**(7043):755-756.
30. Geraci JR, Anderson, D.M., Timperi, R.J., St Aubin, D.J., Early, G.A., Prescott, J.H., Mayo, C.A.: **Humpback whales (*Megaoetere novaeangliae*) fatally poisoned by dinoflagellate toxin.** *Can J Fish Aq Sc* 1989, **46**:1895-1898.
31. Van Dolah FM: **Diversity of Marine and Freshwater Algal Toxins.** New York: Marcel Dekker; 2000.
32. Larsson U, Elmgren, R., and Wulff, F. : **Eutrophication and the Baltic Sea—Causes and Consequences.** *AMBIO* 1985, **14**:9-14.
33. Oshima Y, Itakura, H., Lee, K.-C., Yasumoto, T., Blackburn, S. and Hallegraeff, G. : **Toxin production by the dinoflagellate *Gymnodinium catenatum*.** New York: Elsevier; 1993.
34. Sako Y, Yoshida, T., Uchida, A., Arakawa, O., Noguchi, T. and Ishida, Y.: **Purification and characterization of a sulfotransferase specific to N-21 of saxitoxin and gonyautoxin 2 + 3 from the toxic dinoflagellate *Gymnodinium catenatum* (Dinophyceae).** *J Phycol* 2001, **37**:1044- 1051.

35. Usup G, Kulis, D. M. and Anderson, D. M. : **Growth and toxin production of the toxic dinoflagellate *Pyrodinium bahamense* var. *compressum* in laboratory cultures.** *Natural Toxins* 1994, **2**:254–262.
36. Van Dolah FM: **Marine algal toxins: origins, health effects, and their increased occurrence.** *Environmental Health Perspectives* 2000, **108**(Supplement 1):133–141.
37. Lehane LaL, R. J.: **Ciguatera: recent advances but the risk remains.** *International Journal of Food Microbiology* 2000, **61**(91–125).
38. James KJ, Moroney C, Roden C, Satake M, Yasumoto T, Lehane M, Furey A: **Ubiquitous 'benign' alga emerges as the cause of shellfish contamination responsible for the human toxic syndrome, azaspiracid poisoning.** *Toxicon* 2003, **41**(2):145-151.
39. Andersson JO: **Lateral gene transfer in eukaryotes.** *Cell Mol Life Sci* 2005, **62**(11):1182-1197.
40. Archibald JM, Rogers MB, Toop M, Ishida K, Keeling PJ: **Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga *Bigeloviella natans*.** *Proc Natl Acad Sci U S A* 2003, **100**(13):7678-7683.
41. Bergthorsson U, Adams KL, Thomason B, Palmer JD: **Widespread horizontal transfer of mitochondrial genes in flowering plants.** *Nature* 2003, **424**(6945):197-201.
42. Loftus B, Anderson I, Davies R, Alsmark UC, Samuelson J, Amedeo P, Roncaglia P, Berriman M, Hirt RP, Mann BJ *et al*: **The genome of the protist parasite *Entamoeba histolytica*.** *Nature* 2005, **433**(7028):865-868.
43. Stiller JW: **Experimental design and statistical rigor in phylogenomics of horizontal and endosymbiotic gene transfer.** *BMC Evol Biol*, **11**:259.
44. Lapierre P, Lasek-Nesselquist E, Gogarten JP: **The impact of HGT on phylogenomic reconstruction methods.** *Brief Bioinform.*
45. Nosenko T, Bhattacharya D: **Horizontal gene transfer in chromalveolates.** *BMC Evol Biol* 2007, **7**:173.
46. Wisecaver JH, Hackett JD: **Dinoflagellate genome evolution.** *Annu Rev Microbiol* 2011, **65**:369-387.
47. Slamovits CH, Okamoto N, Burri L, James ER, Keeling PJ: **A bacterial proteorhodopsin proton pump in marine eukaryotes.** *Nat Commun*, **2**:183.
48. Herring PJaW, E.A.: **Bioluminescence in the plankton and nekton**, vol. 1. San Diego: Academic Press; 2001.
49. Latz MI, Bovard M, VanDelinder V, Segre E, Rohr J, Groisman A: **Bioluminescent response of individual dinoflagellate cells to hydrodynamic stress measured with millisecond resolution in a microfluidic device.** *J Exp Biol* 2008, **211**(Pt 17):2865-2875.
50. Baker A, Robbins, I., Moline, M.A. and Iglesias-Rodriguez, M.D.: **Oligonucleotide primers for the detection of bioluminescent dinoflagellates reveal novel luciferase sequences and information on the molecular evolution of this gene.** *J Phycol* 2008, **44**(2):419-428.
51. Sweeney BM: **Bioluminescence and circadian rhythms.** Oxford, UK: Blackwell; 1987.

52. White HH: **Effects of dinoflagellate bioluminescence on the ingestion rates of herbivorous zooplankton.** *J Exp Mar Biol Ecol* 1979, **36**:217–224.
53. Esaias WE, and Curl, H. C.: **Effect of dinoflagellate bioluminescence on copepod ingestion rates.** *Limnol Oceanogr* 1972, **17**:901–905.
54. Buskey EJ, Reynolds, G. T., Swift, E. and Walton, A. J.: **Interactions between copepods and bioluminescent dinoflagellates: direct observations using image intensification.** *Biol Bull* 1985, **169**:530.
55. Mensinger AF, and Case, J. F. : **Dinoflagellate luminescence increases susceptibility of zooplankton to teleost predation.** *Mar Biol* 1992, **112**:207–210.
56. Abrahams MVaT, .L. D.: **Bioluminescence in dinoflagellates: a test of the burglar alarm hypothesis.** *Ecology* 1993, **74**:258–260.
57. Fleisher KJaC, J. F.: **Cephalopod predation facilitated by dinoflagellate luminescence.** *Biol Bull* 1995, **189**:263–271.
58. Schmitter RE: **The fine structure of Gonyaulax polyedra, a bioluminescent marine dinoflagellate.** *J Cell Sci* 1971, **9**(1):147-173.
59. Morales-Ramírez A, Viquez, R., Rodríguez, K. and Vargas, M. : **Red tide bloom produced by Lingulodinium polyedrum (Peridinales, Dinophyceae) in Bahía Culebra, Papagayo Gulf, Costa Rica.** *Rev Biol Trop* 2001, **49**(Suppl.2):19-23.
60. Paz B, Daranas AH, Norte M, Riobo P, Franco JM, Fernandez JJ: **Yessotoxins, a group of marine polyether toxins: an overview.** *Mar Drugs* 2008, **6**(2):73-102.
61. Spector DL: **Dinoflagellate nuclei, p. 107–147.** In D L Spector(ed), *Dinoflagellates Academic Press, London, United Kingdom* 1984.
62. Nassoury N, Hastings, J.W. and Morse, D.: **Biochemistry and circadian regulation of output from the Gonyaulax clock: Are there many clocks or simply many hands?** . Georgetown, TX: Landes Bioscience/Eurekah; 2005.
63. Hastings JW: **The biology of circadian rhythms from man to micro-organism.** *N Engl J Med* 1970, **282**(8):435-441.
64. Dunlap JC: **Molecular bases for circadian clocks.** *Cell* 1999, **96**(2):271-290.
65. Dunlap JC, Loros JJ, Liu Y, Crosthwaite SK: **Eukaryotic circadian systems: cycles in common.** *Genes Cells* 1999, **4**(1):1-10.
66. Devlin PF: **Signs of the time: environmental input to the circadian clock.** *J Exp Bot* 2002, **53**(374):1535-1550.
67. Millar AJ: **Input signals to the plant circadian clock.** *J Exp Bot* 2004, **55**(395):277-283.
68. Kondo T, Ishiura M: **The circadian clock of cyanobacteria.** *Bioessays* 2000, **22**(1):10-15.
69. Dunlap J: **Circadian rhythms. An end in the beginning.** *Science* 1998, **280**(5369):1548-1549.
70. Young MW, Kay SA: **Time zones: a comparative genetics of circadian clocks.** *Nat Rev Genet* 2001, **2**(9):702-715.
71. Hastings JW: **The Gonyaulax clock at 50: translational control of circadian expression.** *Cold Spring Harb Symp Quant Biol* 2007, **72**:141-144.
72. Hastings JW, Sweeney BM: **On the Mechanism of Temperature Independence in a Biological Clock.** *Proc Natl Acad Sci U S A* 1957, **43**(9):804-811.

73. Hastings JW, Sweeney BM: **A persistent diurnal rhythm of luminescence in *Gonyaulax polyedra*.** *Biol Bull* 1958, **115**(440-458).
74. Hastings JW, Sweeney BM: **The action spectrum for shifting the phase of the rhythm of luminescence in *Gonyaulax polyedra*.** *J Gen Physiol* 1960, **43**:697-706.
75. Hastings JW: **Fifty years of fun.** *J Biol Rhythms* 2001, **16**(1):5-18.
76. Akimoto H, Wu C, Kinumi T, Ohmiya Y: **Biological rhythmicity in expressed proteins of the marine dinoflagellate *Lingulodinium polyedrum* demonstrated by chronological proteomics.** *Biochem Biophys Res Commun* 2004, **315**(2):306-312.
77. Morse D, Milos PM, Roux E, Hastings JW: **Circadian regulation of bioluminescence in *Gonyaulax* involves translational control.** *Proc Natl Acad Sci U S A* 1989, **86**(1):172-176.
78. Dunlap JC, Hastings JW: **The biological clock in *Gonyaulax* controls luciferase activity by regulating turnover.** *J Biol Chem* 1981, **256**(20):10509-10518.
79. Morse DS, Fritz L, Hastings JW: **What is the clock? Translational regulation of circadian bioluminescence.** *Trends Biochem Sci* 1990, **15**(7):262-265.
80. Colepicolo P, Roenneberg T., Morse, D., Taylor, W.R. and Hastings, J.W.: **Circadian regulation of bioluminescence in the dinoflagellate *Pyrocystis lunula*.** *J Phycol* 1993, **29**(2):173-179.
81. Kurosawa G, Aihara K, Iwasa Y: **A model for the circadian rhythm of cyanobacteria that maintains oscillation without gene expression.** *Biophys J* 2006, **91**(6):2015-2023.
82. Sugano S, Andronis C, Ong MS, Green RM, Tobin EM: **The protein kinase CK2 is involved in regulation of circadian rhythms in *Arabidopsis*.** *Proc Natl Acad Sci U S A* 1999, **96**(22):12362-12366.
83. Daniel X, Sugano S, Tobin EM: **CK2 phosphorylation of CCA1 is necessary for its circadian oscillator function in *Arabidopsis*.** *Proc Natl Acad Sci U S A* 2004, **101**(9):3292-3297.
84. Liu Y, Loros J, Dunlap JC: **Phosphorylation of the *Neurospora* clock protein FREQUENCY determines its degradation rate and strongly influences the period length of the circadian clock.** *Proc Natl Acad Sci U S A* 2000, **97**(1):234-239.
85. Fang Y, Sathyanarayanan S, Sehgal A: **Post-translational regulation of the *Drosophila* circadian clock requires protein phosphatase 1 (PP1).** *Genes Dev* 2007, **21**(12):1506-1518.
86. Vanselow K, Kramer A: **Role of phosphorylation in the mammalian circadian clock.** *Cold Spring Harb Symp Quant Biol* 2007, **72**:167-176.
87. Eide EJ, Kang H, Crapo S, Gallego M, Virshup DM: **Casein kinase I in the mammalian circadian clock.** *Methods Enzymol* 2005, **393**:408-418.
88. Allada R, Meissner RA: **Casein kinase 2, circadian clocks, and the flight from mutagenic light.** *Mol Cell Biochem* 2005, **274**(1-2):141-149.
89. Lin JM, Kilman VL, Keegan K, Paddock B, Emery-Le M, Rosbash M, Allada R: **A role for casein kinase 2alpha in the *Drosophila* circadian clock.** *Nature* 2002, **420**(6917):816-820.

90. Tamaru T, Hirayama J, Isojima Y, Nagai K, Norioka S, Takamatsu K, Sassone-Corsi P: **CK2alpha phosphorylates BMAL1 to regulate the mammalian clock.** *Nat Struct Mol Biol* 2009, **16**(4):446-448.
91. Tsuchiya Y, Akashi M, Matsuda M, Goto K, Miyata Y, Node K, Nishida E: **Involvement of the protein kinase CK2 in the regulation of mammalian circadian rhythms.** *Sci Signal* 2009, **2**(73):ra26.
92. Yang Y, Cheng P, Liu Y: **Regulation of the Neurospora circadian clock by casein kinase II.** *Genes Dev* 2002, **16**(8):994-1006.
93. Comolli JC, Hastings JW: **Novel effects on the Gonyaulax circadian system produced by the protein kinase inhibitor staurosporine.** *J Biol Rhythms* 1999, **14**(1):11-19.
94. Comolli J, Taylor W, Hastings JW: **An inhibitor of protein phosphorylation stops the circadian oscillator and blocks light-induced phase shifting in Gonyaulax polyedra.** *J Biol Rhythms* 1994, **9**(1):13-26.
95. Comolli J, Taylor W, Rehman J, Hastings JW: **Inhibitors of serine/threonine phosphoprotein phosphatases alter circadian properties in Gonyaulax polyedra.** *Plant Physiol* 1996, **111**(1):285-291.
96. Comolli JC, Fagan T, Hastings JW: **A type-1 phosphoprotein phosphatase from a dinoflagellate as a possible component of the circadian mechanism.** *J Biol Rhythms* 2003, **18**(5):367-376.
97. LaJeunesse TC, Lambert, G., Andersen, A., Coffroth, M.A., Galbraith, D.W. : **Symbiodinium (Pyrrhophyta) genome sizes (DNA content) are smallest among dinoflagellates.** *J Phycol* 2005, **41**:880-886.
98. Veldhuis MJW, Cucci, T.L., Sieracki, M.E.: **Cellular DNA content of marine phytoplankton using two new fluorochromes: taxonomic and ecological implications.** *J Phycol* 1997, **33**:527-541.
99. Lynch M, Conery JS: **The origins of genome complexity.** *Science* 2003, **302**(5649):1401-1404.
100. Gregory TR: **Synergy between sequence and size in large-scale genomics.** *Nat Rev Genet* 2005, **6**(9):699-708.
101. Hou Y, Lin S: **Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes.** *PLoS One* 2009, **4**(9):e6978.
102. Bayer T, Aranda M, Sunagawa S, Yum LK, Desalvo MK, Lindquist E, Coffroth MA, Voolstra CR, Medina M: **Symbiodinium transcriptomes: genome insights into the dinoflagellate symbionts of reef-building corals.** *PLoS One* 2012, **7**(4):e35269.
103. Rizzo PJ: **The enigma of the dinoflagellate chromosome.** *Journal of Protozoology* 1991, **38**:246-252.
104. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ: **Crystal structure of the nucleosome core particle at 2.8 Å resolution.** *Nature* 1997, **389**(6648):251-260.
105. Arents G, Burlingame RW, Wang BC, Love WE, Moudrianakis EN: **The nucleosomal core histone octamer at 3.1 Å resolution: a tripartite protein assembly and a left-handed superhelix.** *Proc Natl Acad Sci U S A* 1991, **88**(22):10148-10152.

106. Luger K, Richmond TJ: **DNA binding within the nucleosome core.** *Curr Opin Struct Biol* 1998, **8**(1):33-40.
107. Davey CA, Sargent DF, Luger K, Maeder AW, Richmond TJ: **Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution.** *J Mol Biol* 2002, **319**(5):1097-1113.
108. Ruthenburg AJ, Li H, Patel DJ, Allis CD: **Multivalent engagement of chromatin modifications by linked binding modules.** *Nat Rev Mol Cell Biol* 2007, **8**(12):983-994.
109. Thiriet C, Hayes JJ: **Replication-independent core histone dynamics at transcriptionally active loci in vivo.** *Genes Dev* 2005, **19**(6):677-682.
110. Thambirajah AA, Dryhurst D, Ishibashi T, Li A, Maffey AH, Ausio J: **H2A.Z stabilizes chromatin in a way that is dependent on core histone acetylation.** *J Biol Chem* 2006, **281**(29):20036-20044.
111. Rizzo PJ, Burghardt RC: **Histone-like protein and chromatin structure in the wall-less dinoflagellate *Gymnodinium nelsoni*.** *Biosystems* 1982, **15**(1):27-34.
112. Wong JT, New DC, Wong JC, Hung VK: **Histone-like proteins of the dinoflagellate *Cryptocodinium cohnii* have homologies to bacterial DNA-binding proteins.** *Eukaryot Cell* 2003, **2**(3):646-650.
113. Chan YH, Wong JT: **Concentration-dependent organization of DNA by the dinoflagellate histone-like protein HCC3.** *Nucleic Acids Res* 2007, **35**(8):2573-2583.
114. Geraud M-L, Sala-Rovira, M., Herzog, M. and Soyer-Gobillard, M-O.: **Immunocytochemical localization of the DNA-binding protein HCC during the cell cycle of the histone-less dinoflagellate protist *Cryptocodinium cohnii* b.** *Biol Cell* 1991, **71**(123-134).
115. Chudnovsky Y, Li JF, Rizzo PJ, Hastings JW, Fagan T: **Cloning, Expression, and Characterization of a histone-like protein from the marine dinoflagellate *Lingulodinium polyedrum*.** *Journal of Phycology* 2002, **38**(3):543-550.
116. Chow MH, Yan KT, Bennett MJ, Wong JT: **Birefringence and DNA condensation of liquid crystalline chromosomes.** *Eukaryot Cell* 2010, **9**(10):1577-1587.
117. Lin S: **Genomic understanding of dinoflagellates.** *Res Microbiol* 2011.
118. Suzuki MM, Bird A: **DNA methylation landscapes: provocative insights from epigenomics.** *Nat Rev Genet* 2008, **9**(6):465-476.
119. Miranda TB, Jones PA: **DNA methylation: the nuts and bolts of repression.** *J Cell Physiol* 2007, **213**(2):384-390.
120. Steele RE, Rae PM: **Ordered distribution of modified bases in the DNA of a dinoflagellate.** *Nucleic Acids Res* 1980, **8**(20):4709-4725.
121. Rae PM: **Hydroxymethyluracil in eukaryote DNA: a natural feature of the pyrrrophyta (dinoflagellates).** *Science* 1976, **194**(4269):1062-1064.
122. Dale B: **Dinoflagellate resting cysts: 'benthic plankton'.** Cambridge: Cambridge University Press; 1983.
123. Von Stosch HA: **Observation on the vegetative reproduction and sexual life cycles of two freshwater dinoflagellates, *Gymnodinium pseudopalustre* Schiller and *Woloszynskia apiculata* sp. nov.** *Br Phycol J* 1973, **8**:105-134.

124. Lewis J, Harris, A., Jones, K. and Edmonds, R.: **Long-term survival of marine planktonic diatoms and dinoflagellates in stored sediment samples.** *J Plankton Res* 1999, **21**:343–354.
125. Matsuoka KaF, Y.: **Taxonomy of cysts.** In: *Manual on Harmful Marine Microalgae.* Edited by Hallegraeff GM, Anderson, D. M. and Cembella, A. D., vol. No. 33. Paris: IOC Manuals and Guides; 1995 381–401.
126. Schmitter RE: **Temporary cysts of *Gonyaulax excavata*: effects of temperature and light.** New York: Elsevier North Holland; 1979.
127. Tsim ST, Wong JT, Wong YH: **Calcium ion dependency and the role of inositol phosphates in melatonin-induced encystment of dinoflagellates.** *J Cell Sci* 1997, **110 (Pt 12)**:1387-1393.
128. Jensen MO, Moestrup, O. : **Autecology of the toxic dinoflagellate *Alexandrium ostenfeldii*: Life history and growth at different temperatures and salinities.** *Eur J Phycol* 1997, **32**:9-18.
129. Nagasaki K, Yamaguchi, M., Imai, I.: **Algicidal activity of a killer bacterium against the harmful red tide dinoflagellate *Heterocapsa circularisquama* isolated from Ago Bay, Japan.** . *Nippon-Suisan-Gakkaish* 2000, **66**:666–673.
130. Kita T, Fukuyo, Y., Tokuda, H., Hirano, R. : **Life history and ecology of *Goniodoma pseudogoniaulax* (Pyrrhophyta) in a rock pool.** *Bulletin of Marine Sciences* 1985, **37**:643 - 651.
131. Field CB, Behrenfeld MJ, Randerson JT, Falkowski P: **Primary production of the biosphere: integrating terrestrial and oceanic components.** *Science* 1998, **281**(5374):237-240.
132. Muscatine L, McCloskey, L.R., Marian, R.E.: **Estimating the daily contribution of carbon from zooxanthellae to coral animal respiration.** *Limnology and Oceanography* 1981, **26**:601- 611.
133. Camacho FG, Rodriguez JG, Miron AS, Garcia MC, Belarbi EH, Chisti Y, Grima EM: **Biotechnological significance of toxic marine dinoflagellates.** *Biotechnol Adv* 2007, **25**(2):176-194.
134. Schmitter RE, Njus D, Sulzman FM, Gooch VD, Hastings JW: **Dinoflagellate bioluminescence: a comparative study of invitro components.** *J Cell Physiol* 1976, **87**(1):123-134.
135. Hastings JW, and Sweeney, B. M.: **A persistent diurnal rhythm of luminescence in *Gonyaulax polyedra*.** *Biological Bulletin* 1958, **115**:444-458.
136. Hastings JW, Astrachan L, Sweeney BM: **A persistent daily rhythm in photosynthesis.** *J Gen Physiol* 1961, **45**:69-76.
137. Sweeney BM: **The photosynthetic rhythm in single cells of *Gonyaulax polyedra*.** *Cold Spring Harb Symp Quant Biol* 1960, **25**:145-148.
138. Roenneberg T, Colfax GN, Hastings JW: **A circadian rhythm of population behavior in *Gonyaulax polyedra*.** *J Biol Rhythms* 1989, **4**(2):201-216.
139. Livolant F: **Cholesteric organization of DNA in vivo and in vitro.** *Eur J Cell Biol* 1984, **33**(2):300-311.
140. Livolant F: **Positive and negative birefringence in chromosomes.** *Chromosoma* 1978, **68**(1):45-58.

141. Herzog M, Soyer MO: **The native structure of dinoflagellate chromosomes and their stabilization by Ca²⁺ and Mg²⁺ cations.** *Eur J Cell Biol* 1983, **30**(1):33-41.
142. Sigeo DC: **Structural DNA and genetically active DNA in dinoflagellate chromosomes.** *Biosystems* 1983, **16**(3-4):203-210.
143. Kornberg RD: **The molecular basis of eukaryotic transcription.** *Proc Natl Acad Sci U S A* 2007, **104**(32):12955-12961.
144. Smale ST, Kadonaga JT: **The RNA polymerase II core promoter.** *Annu Rev Biochem* 2003, **72**:449-479.
145. Hahn S: **Structure and mechanism of the RNA polymerase II transcription machinery.** *Nat Struct Mol Biol* 2004, **11**(5):394-403.
146. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC *et al*: **Genome-wide analysis of mammalian promoter architecture and evolution.** *Nat Genet* 2006, **38**(6):626-635.
147. Everett RD, Baty D, Chambon P: **The repeated GC-rich motifs upstream from the TATA box are important elements of the SV40 early promoter.** *Nucleic Acids Res* 1983, **11**(8):2447-2464.
148. Yoshikawa T, Takishita, K., Ishida, Y., Uchida, A.: **Molecular cloning and nucleotide sequence analysis of the gene coding for chloroplast-type ferredoxin from the dinoflagellates *Peridinium bipes* and *Alexandrium tamarense*.** *Fisheries Science* 1997, **63**(5):692-700.
149. Wong JM, Liu F, Bateman E: **Isolation of genomic DNA encoding transcription factor TFIID from *Acanthamoeba castellanii*: characterization of the promoter.** *Nucleic Acids Res* 1992, **20**(18):4817-4824.
150. Huang W, Bateman E: **Cloning, expression, and characterization of the TATA-binding protein (TBP) promoter binding factor, a transcription activator of the *Acanthamoeba* TBP gene.** *J Biol Chem* 1995, **270**(48):28839-28847.
151. Cohen SM, Knecht D, Lodish HF, Loomis WF: **DNA sequences required for expression of a *Dictyostelium* actin gene.** *EMBO J* 1986, **5**(12):3361-3366.
152. Kimmel AR, Firtel RA: **Sequence organization in *Dictyostelium*: unique structure at the 5'-ends of protein coding genes.** *Nucleic Acids Res* 1983, **11**(2):541-552.
153. Liston DR, Johnson PJ: **Analysis of a ubiquitous promoter element in a primitive eukaryote: early evolution of the initiator element.** *Mol Cell Biol* 1999, **19**(3):2380-2388.
154. McAndrew MB, Read M, Sims PF, Hyde JE: **Characterisation of the gene encoding an unusually divergent TATA-binding protein (TBP) from the extremely A+T-rich human malaria parasite *Plasmodium falciparum*.** *Gene* 1993, **124**(2):165-171.
155. Luo H, Gilinger G, Mukherjee D, Bellofatto V: **Transcription initiation at the TATA-less spliced leader RNA gene promoter requires at least two DNA-binding proteins and a tripartite architecture that includes an initiator element.** *J Biol Chem* 1999, **274**(45):31947-31954.
156. Quon DV, Delgadillo MG, Johnson PJ: **Transcription in the early diverging eukaryote *Trichomonas vaginalis*: an unusual RNA polymerase II and alpha-amanitin-resistant transcription of protein-coding genes.** *J Mol Evol* 1996, **43**(3):253-262.

157. Quon DV, Delgadillo MG, Khachi A, Smale ST, Johnson PJ: **Similarity between a ubiquitous promoter element in an ancient eukaryote and mammalian initiator elements.** *Proc Natl Acad Sci U S A* 1994, **91**(10):4579-4583.
158. Li L, Hastings JW: **The structure and organization of the luciferase gene in the photosynthetic dinoflagellate *Gonyaulax polyedra*.** *Plant Mol Biol* 1998, **36**(2):275-284.
159. Le QH, Markovic P, Hastings JW, Jovine RV, Morse D: **Structure and organization of the peridinin-chlorophyll a-binding protein gene in *Gonyaulax polyedra*.** *Mol Gen Genet* 1997, **255**(6):595-604.
160. Machabee S, Wall L, Morse D: **Expression and genomic organization of a dinoflagellate gene family.** *Plant Mol Biol* 1994, **25**(1):23-31.
161. Lee DH, Mittag M, Sczekan S, Morse D, Hastings JW: **Molecular cloning and genomic organization of a gene for luciferin-binding protein from the dinoflagellate *Gonyaulax polyedra*.** *The Journal of biological chemistry* 1993, **268**(12):8842-8850.
162. Bachvaroff TR, Place AR: **From stop to start: tandem gene arrangement, copy number and trans-splicing sites in the dinoflagellate *Amphidinium carterae*.** *PLoS One* 2008, **3**(8):e2929.
163. Okamoto OK, Liu L, Robertson DL, Hastings JW: **Members of a dinoflagellate luciferase gene family differ in synonymous substitution rates.** *Biochemistry* 2001, **40**(51):15862-15868.
164. Jackson AP: **Tandem gene arrays in *Trypanosoma brucei*: comparative phylogenomic analysis of duplicate sequence variation.** *BMC Evol Biol* 2007, **7**:54.
165. Beauchemin M, Roy S, Daoust P, Dagenais-Bellefeuille S, Bertomeu T, Letourneau L, Lang BF, Morse D: **Dinoflagellate tandem array gene transcripts are highly conserved and not polycistronic.** *Proc Natl Acad Sci U S A* 2012, **109**(39):15793-15798.
166. Rizzo PJ: **RNA synthesis in isolated nuclei of the dinoflagellate *Cryptocodinium cohnii*.** *J Protozool* 1979, **26**(2):290-294.
167. Palenchar JB, Bellofatto V: **Gene transcription in trypanosomes.** *Mol Biochem Parasitol* 2006, **146**(2):135-141.
168. Orphanides G, Lagrange T, Reinberg D: **The general transcription factors of RNA polymerase II.** *Genes Dev* 1996, **10**(21):2657-2683.
169. Conaway JW, Bond MW, Conaway RC: **An RNA polymerase II transcription system from rat liver. Purification of an essential component.** *J Biol Chem* 1987, **262**(17):8293-8297.
170. Conaway RC, Conaway JW: **An RNA polymerase II transcription factor has an associated DNA-dependent ATPase (dATPase) activity strongly stimulated by the TATA region of promoters.** *Proc Natl Acad Sci U S A* 1989, **86**(19):7356-7360.
171. Conaway JW, Conaway RC: **A multisubunit transcription factor essential for accurate initiation by RNA polymerase II.** *J Biol Chem* 1989, **264**(4):2357-2362.
172. Conaway JW, Reines D, Conaway RC: **Transcription initiated by RNA polymerase II and purified transcription factors from liver. Cooperative action**

- of transcription factors tau and epsilon in initial complex formation. *J Biol Chem* 1990, **265**(13):7552-7558.**
173. Sumimoto H, Ohkuma Y, Yamamoto T, Horikoshi M, Roeder RG: **Factors involved in specific transcription by mammalian RNA polymerase II: identification of general transcription factor TFIIG.** *Proc Natl Acad Sci U S A* 1990, **87**(23):9158-9162.
 174. Poon D, Bai Y, Campbell AM, Bjorklund S, Kim YJ, Zhou S, Kornberg RD, Weil PA: **Identification and characterization of a TFIID-like multiprotein complex from *Saccharomyces cerevisiae*.** *Proc Natl Acad Sci U S A* 1995, **92**(18):8224-8228.
 175. Sanders SL, Weil PA: **Identification of two novel TAF subunits of the yeast *Saccharomyces cerevisiae* TFIID complex.** *J Biol Chem* 2000, **275**(18):13895-13900.
 176. Chatterjee S, Struhl K: **Connecting a promoter-bound protein to TBP bypasses the need for a transcriptional activation domain.** *Nature* 1995, **374**(6525):820-822.
 177. Chong JA, Moran MM, Teichmann M, Kaczmarek JS, Roeder R, Clapham DE: **TATA-binding protein (TBP)-like factor (TLF) is a functional regulator of transcription: reciprocal regulation of the neurofibromatosis type 1 and c-fos genes by TLF/TRF2 and TBP.** *Mol Cell Biol* 2005, **25**(7):2632-2643.
 178. Holmes MC, Tjian R: **Promoter-selective properties of the TBP-related factor TRF1.** *Science* 2000, **288**(5467):867-870.
 179. Guillebault D, Sasorith S, Derelle E, Wurtz JM, Lozano JC, Bingham S, Tora L, Moreau H: **A new class of transcription initiation factors, intermediate between TATA box-binding proteins (TBPs) and TBP-like factors (TLFs), is present in the marine unicellular organism, the dinoflagellate *Cryptothecodinium cohnii*.** *J Biol Chem* 2002, **277**(43):40881-40886.
 180. Toulza E, Shin MS, Blanc G, Audic S, Laabir M, Collos Y, Claverie JM, Grzebyk D: **Gene expression in proliferating cells of the dinoflagellate *Alexandrium catenella* (Dinophyceae).** *Appl Environ Microbiol* 2010, **76**(13):4521-4529.
 181. Qiu XB, Lin YL, Thome KC, Pian P, Schlegel BP, Weremowicz S, Parvin JD, Dutta A: **An eukaryotic RuvB-like protein (RUVBL1) essential for growth.** *J Biol Chem* 1998, **273**(43):27786-27793.
 182. Eickbush TH, Moudrianakis EN: **The histone core complex: an octamer assembled by two sets of protein-protein interactions.** *Biochemistry* 1978, **17**(23):4955-4964.
 183. Kasinsky HE, Lewis JD, Dacks JB, Ausio J: **Origin of H1 linker histones.** *FASEB J* 2001, **15**(1):34-42.
 184. Rizzo PJ: **Those amazing dinoflagellate chromosomes.** *Cell Res* 2003, **13**(4):215-217.
 185. Vernet G, Sala-Rovira M, Maeder M, Jacques F, Herzog M: **Basic nuclear proteins of the histone-less eukaryote *Cryptothecodinium cohnii* (Pyrrhophyta): two-dimensional electrophoresis and DNA-binding properties.** *Biochim Biophys Acta* 1990, **1048**(2-3):281-289.
 186. Rizzo PJ: **Comparative aspects of basic chromatin proteins in dinoflagellates.** *Biosystems* 1981, **14**(3-4):433-443.

187. Bodansky S, Mintz LB, Holmes DS: **The mesokaryote Gyrodinium cohnii lacks nucleosomes.** *Biochem Biophys Res Commun* 1979, **88**(4):1329-1336.
188. Rizzo PJ, Nooden LD: **Chromosomal proteins in the dinoflagellate alga Gyrodinium cohnii.** *Science* 1972, **176**(36):796-797.
189. Roy S, Morse D: **A full suite of histone and histone modifying genes are transcribed in the dinoflagellate lingulodinium.** *PLoS One* 2012, **7**(3):e34340.
190. Livolant F: **Cholesteric organization of DNA in the stallion sperm head.** *Tissue Cell* 1984, **16**(4):535-555.
191. Balhorn R: **The protamine family of sperm nuclear proteins.** *Genome Biol* 2007, **8**(9):227.
192. Rizzo PJ, Jones M, Ray SM: **Isolation and properties of isolated nuclei from the Florida red tide dinoflagellate Gymnodinium breve (Davis).** *J Protozool* 1982, **29**(2):217-222.
193. Kellenberger E, Arnold-Schulz-Gahmen B: **Chromatins of low-protein content: special features of their compaction and condensation.** *FEMS Microbiol Lett* 1992, **79**(1-3):361-370.
194. Holck A, Lossius I, Aasland R, Haarr L, Kleppe K: **DNA- and RNA-binding proteins of chromatin from Escherichia coli.** *Biochim Biophys Acta* 1987, **908**(2):188-199.
195. Sala-Rovira M, Geraud ML, Caput D, Jacques F, Soyer-Gobillard MO, Vernet G, Herzog M: **Molecular cloning and immunolocalization of two variants of the major basic nuclear protein (HCC) from the histone-less eukaryote Crypthecodinium cohnii (Pyrrhophyta).** *Chromosoma* 1991, **100**(8):510-518.
196. Gornik SG, Ford KL, Mulhern TD, Bacic A, McFadden GI, Waller RF: **Loss of nucleosomal DNA condensation coincides with appearance of a novel nuclear protein in dinoflagellates.** *Curr Biol*, **22**(24):2303-2312.
197. Azevedo C: **Fine structure of Perkinsus atlanticus n. sp. (Apicomplexa, Perkinsea) parasite of the clam Ruditapes decussatus from Portugal.** *J Parasitol* 1989, **75**(4):627-635.
198. Jaeckisch N, Yang I, Wohlrab S, Glockner G, Kroymann J, Vogel H, Cembella A, John U: **Comparative genomic and transcriptomic characterization of the toxigenic marine dinoflagellate Alexandrium ostenfeldii.** *PLoS One* 2011, **6**(12):e28012.
199. Minguez A, Franca S, Moreno Diaz de la Espina S: **Dinoflagellates have a eukaryotic nuclear matrix with lamin-like proteins and topoisomerase II.** *J Cell Sci* 1994, **107** (Pt 10):2861-2873.
200. Dechat T, Pfliegerhaer K, Sengupta K, Shimi T, Shumaker DK, Solimando L, Goldman RD: **Nuclear lamins: major factors in the structural organization and function of the nucleus and chromatin.** *Genes Dev* 2008, **22**(7):832-853.
201. Dechat T, Adam SA, Goldman RD: **Nuclear lamins and chromatin: when structure meets function.** *Adv Enzyme Regul* 2009, **49**(1):157-166.
202. Dechat T, Adam SA, Taimen P, Shimi T, Goldman RD: **Nuclear lamins.** *Cold Spring Harb Perspect Biol* 2010, **2**(11):a000547.
203. Bhaud Y, Geraud ML, Ausseil J, Soyer-Gobillard MO, Moreau H: **Cyclic expression of a nuclear protein in a dinoflagellate.** *J Eukaryot Microbiol* 1999, **46**(3):259-267.

204. Guillebault D, Derelle E, Bhaud Y, Moreau H: **Role of nuclear WW domains and proline-rich proteins in dinoflagellate transcription.** *Protist* 2001, **152**(2):127-138.
205. Boggon TJ, Shan WS, Santagata S, Myers SC, Shapiro L: **Implication of tubby proteins as transcription factors by structure-based functional analysis.** *Science* 1999, **286**(5447):2119-2125.
206. Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA: **Structure and evolution of transcriptional regulatory networks.** *Curr Opin Struct Biol* 2004, **14**(3):283-291.
207. Sommerville J: **Activities of cold-shock domain proteins in translation control.** *Bioessays* 1999, **21**(4):319-325.
208. Balaji S, Babu MM, Iyer LM, Aravind L: **Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains.** *Nucleic Acids Res* 2005, **33**(13):3994-4006.
209. Bird A: **DNA methylation patterns and epigenetic memory.** *Genes Dev* 2002, **16**(1):6-21.
210. Blank RJ, Huss, V.A.R., Kersten, W.: **Base composition of DNA from symbiotic dinoflagellates: a tool for phylogenetic classification.** *Arch Microbiol* 1988, **149**:515-520.
211. ten Lohuis MR, Miller DJ: **Light-regulated transcription of genes encoding peridinin chlorophyll a proteins and the major intrinsic light-harvesting complex proteins in the dinoflagellate amphidinium carterae hulburt (Dinophyceae). Changes In cytosine methylation accompany photoadaptation.** *Plant Physiol* 1998, **117**(1):189-196.
212. Rae PM, Steele RE: **Modified bases in the DNAs of unicellular eukaryotes: an examination of distributions and possible roles, with emphasis on hydroxymethyluracil in dinoflagellates.** *Biosystems* 1978, **10**(1-2):37-53.
213. Teebor GW, Frenkel K, Goldstein MS: **Ionizing radiation and tritium transmutation both cause formation of 5-hydroxymethyl-2'-deoxyuridine in cellular DNA.** *Proc Natl Acad Sci U S A* 1984, **81**(2):318-321.
214. Thomas S, Green A, Sturm NR, Campbell DA, Myler PJ: **Histone acetylations mark origins of polycistronic transcription in Leishmania major.** *BMC Genomics* 2009, **10**:152.
215. Wong JT, Kwok AC: **Proliferation of dinoflagellates: blooming or bleaching.** *Bioessays* 2005, **27**(7):730-740.
216. Gyula P, Schafer E, Nagy F: **Light perception and signalling in higher plants.** *Curr Opin Plant Biol* 2003, **6**(5):446-452.
217. Van Dolah F.M L, K.B., More, J.S., Brunelle, S.A., Ryan, J.C., Monroe, E.A., Haynes, B.L.: **Microarray analysis of diurnal- and circadian-regulated genes in the Florida red-tide dinoflagellate Karenia brevis (Dinophyceae).** *J Phycol* 2007, **43**(4):741-752.
218. Lesser MP: **Elevated temperatures and ultraviolet radiation cause oxidative stress and inhibit photosynthesis in symbiotic dinoflagellates.** *Limnol Oceanogr* 1996, **41**(2):271-283.

219. Lesser MP: **Oxidative stress causes coral bleaching during exposure to elevated temperatures.** *Coral Reefs* 1997, **16**:187-192.
220. Rosic NN, Pernice M, Dove S, Dunn S, Hoegh-Guldberg O: **Gene expression profiles of cytosolic heat shock proteins Hsp70 and Hsp90 from symbiotic dinoflagellates in response to thermal stress: possible implications for coral bleaching.** *Cell Stress Chaperones*, **16**(1):69-80.
221. Walsh CT, Garneau-Tsodikova S, Gatto GJ, Jr.: **Protein posttranslational modifications: the chemistry of proteome diversifications.** *Angew Chem Int Ed Engl* 2005, **44**(45):7342-7372.
222. Okamoto OK, Asano, C.S., Aidar, E., Colepiccolo, P.: **of cadmium on growth and superoxide dismutase activity of this species of dinoflagellate. the marine microalga Tetraselmis gracilis.** *J Phycol* 1996, **32**:74-79.
223. Okamoto OK, Colepiccolo P: **Response of superoxide dismutase to pollutant metal stress in the marine dinoflagellate Gonyaulax polyedra.** *Comp Biochem Physiol C Pharmacol Toxicol Endocrinol* 1998, **119**(1):67-73.
224. Okamoto OK, Robertson DL, Fagan TF, Hastings JW, Colepiccolo P: **Different regulatory mechanisms modulate the expression of a dinoflagellate iron-superoxide dismutase.** *J Biol Chem* 2001, **276**(23):19989-19993.
225. Okamoto OK, Hastings JW: **Genome-wide analysis of redox-regulated genes in a dinoflagellate.** *Gene* 2003, **321**:73-81.
226. Guo R, Ebenezer V, Ki JS: **Transcriptional responses of heat shock protein 70 (Hsp70) to thermal, bisphenol A, and copper stresses in the dinoflagellate Prorocentrum minimum.** *Chemosphere*, **89**(5):512-520.
227. Guo R, Ki JS: **Differential transcription of heat shock protein 90 (HSP90) in the dinoflagellate Prorocentrum minimum by copper and endocrine-disrupting chemicals.** *Ecotoxicology*, **21**(5):1448-1457.
228. Lowe CD, Mello LV, Samatar N, Martin LE, Montagnes DJ, Watts PC: **The transcriptome of the novel dinoflagellate Oxyrrhis marina (Alveolata: Dinophyceae): response to salinity examined by 454 sequencing.** *BMC Genomics*, **12**:519.
229. McClung CR: **Plant circadian rhythms.** *Plant Cell* 2006, **18**(4):792-803.
230. Loros JJ, Dunlap JC: **Genetic and molecular analysis of circadian rhythms in Neurospora.** *Annu Rev Physiol* 2001, **63**:757-794.
231. Rivkees SA: **The Development of Circadian Rhythms: From Animals To Humans.** *Sleep Med Clin* 2007, **2**(3):331-341.
232. Roenneberg T, Rehman J: **Nitrate, a nonphotic signal for the circadian system.** *FASEB J* 1996, **10**(12):1443-1447.
233. Roenneberg T, Mrosovsky M: **Entrainment of the human circadian clock.** *Cold Spring Harb Symp Quant Biol* 2007, **72**:293-299.
234. Mrosovsky M, Roenneberg T: **Circadian entrainment of Neurospora crassa.** *Cold Spring Harb Symp Quant Biol* 2007, **72**:279-285.
235. Roenneberg T, Kumar CJ, Mrosovsky M: **The human circadian clock entrains to sun time.** *Curr Biol* 2007, **17**(2):R44-45.
236. Woelfle MA, Johnson CH: **No promoter left behind: global circadian gene expression in cyanobacteria.** *J Biol Rhythms* 2006, **21**(6):419-431.

237. Ito H, Mutsuda M, Murayama Y, Tomita J, Hosokawa N, Terauchi K, Sugita C, Sugita M, Kondo T, Iwasaki H: **Cyanobacterial daily life with Kai-based circadian and diurnal genome-wide transcriptional control in *Synechococcus elongatus***. *Proc Natl Acad Sci U S A* 2009, **106**(33):14168-14173.
238. Okamoto OKaH, J.W.: **Novel dinoflagellate clock-related genes identified through microarray analysis**. *J Phycol* 2003, **39**:519-526.
239. Walz B, Walz, A, Sweeney, B.M. : **A circadian rhythm in RNA in the dinoflagellate, *Gonyaulax polyedra***. *Journal of comparative physiology* 1983, **151**(2):207-213.
240. Dagenais-Bellefeuille S, Bertomeu T, Morse D: **S-phase and M-phase timing are under independent circadian control in the dinoflagellate *Lingulodinium***. *J Biol Rhythms* 2008, **23**(5):400-408.
241. Bertomeu T, Rivoal J, Morse D: **A dinoflagellate CDK5-like cyclin-dependent kinase**. *Biol Cell* 2007, **99**(9):531-540.
242. Karakashian MW, Hastings JW: **The inhibition of a biological clock by actinomycin D**. *Proc Natl Acad Sci U S A* 1962, **48**:2130-2137.
243. Rossini C, Taylor W, Fagan T, Hastings JW: **Lifetimes of mRNAs for clock-regulated proteins in a dinoflagellate**. *Chronobiol Int* 2003, **20**(6):963-976.
244. Morey JS, Monroe EA, Kinney AL, Beal M, Johnson JG, Hitchcock GL, Van Dolah FM: **Transcriptomic response of the red tide dinoflagellate, *Karenia brevis*, to nitrogen and phosphorus depletion and addition**. *BMC Genomics* 2011, **12**:346.
245. Lee TC, Kwok OT, Ho KC, Lee FW: **Effects of different nitrate and phosphate concentrations on the growth and toxin production of an *Alexandrium tamarense* strain collected from Drake Passage**. *Mar Environ Res*, **81**:62-69.
246. Yang I, Beszteri, S., Tillmann, U., Cembella, A., John, U.: **Growth- and nutrient-dependent gene expression in the toxigenic marine dinoflagellate *Alexandrium minutum***. *Harmful Algae* 2011, **12**:55-69.
247. Moustafa A, Evans AN, Kulis DM, Hackett JD, Erdner DL, Anderson DM, Bhattacharya D: **Transcriptome profiling of a toxic dinoflagellate reveals a gene-rich protist and a potential impact on gene expression due to bacterial presence**. *PLoS One* 2010, **5**(3):e9688.
248. Johnson JG, Morey JS, Neely MG, Ryan JC, Van Dolah FM: **Transcriptome remodeling associated with chronological aging in the dinoflagellate, *Karenia brevis***. *Mar Genomics*, **5**:15-25.
249. Yang I, John U, Beszteri S, Glockner G, Krock B, Goesmann A, Cembella AD: **Comparative gene expression in toxic versus non-toxic strains of the marine dinoflagellate *Alexandrium minutum***. *BMC Genomics* 2010, **11**:248.
250. Salcedo T, Upadhyay RJ, Nagasaki K, Bhattacharya D: **Dozens of toxin-related genes are expressed in a nontoxic strain of the dinoflagellate *Heterocapsa circularisquama***. *Mol Biol Evol*, **29**(6):1503-1506.
251. Gast RJ, Beaudoin DJ, Caron DA: **Isolation of symbiotically expressed genes from the dinoflagellate symbiont of the solitary radiolarian *Thalassicolla nucleata***. *Biol Bull* 2003, **204**(2):210-214.

252. Bertucci A, Tambutte E, Tambutte S, Allemand D, Zoccola D: **Symbiosis-dependent gene expression in coral-dinoflagellate association: cloning and characterization of a P-type H⁺-ATPase gene.** *Proc Biol Sci* 2010, **277**(1678):87-95.
253. Leggat W, Seneca F, Wasmund K, Ukani L, Yellowlees D, Ainsworth TD: **Differential responses of the coral host and their algal symbiont to thermal stress.** *PLoS One* 2011, **6**(10):e26687.
254. Yang E, van Nimwegen E, Zavolan M, Rajewsky N, Schroeder M, Magnasco M, Darnell JE, Jr.: **Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes.** *Genome Res* 2003, **13**(8):1863-1872.
255. Kinniburgh AJ, Mertz JE, Ross J: **The precursor of mouse beta-globin messenger RNA contains two intervening RNA sequences.** *Cell* 1978, **14**(3):681-693.
256. Chow LT, Gelinas RE, Broker TR, Roberts RJ: **An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA.** *Cell* 1977, **12**(1):1-8.
257. Berget SM, Moore C, Sharp PA: **Spliced segments at the 5' terminus of adenovirus 2 late mRNA.** *Proc Natl Acad Sci U S A* 1977, **74**(8):3171-3175.
258. Zhang H, Lin, S.: **Complex gene structure of the form II Rubisco in the dinoflagellate *Prorocentrum minimum* (Dinophyceae).** *39* 2003:1160 - 1171.
259. Rowan R, Whitney SM, Fowler A, Yellowlees D: **Rubisco in marine symbiotic dinoflagellates: form II enzymes in eukaryotic oxygenic phototrophs encoded by a nuclear multigene family.** *Plant Cell* 1996, **8**(3):539-553.
260. Orr RJ, Stuken A, Murray SA, Jakobsen KS: **Evolutionary Acquisition and Loss of Saxitoxin Biosynthesis in Dinoflagellates: the Second "Core" Gene, sxtG.** *Appl Environ Microbiol*, **79**(7):2128-2136.
261. Hoppenrath M, Leander BS: **Dinoflagellate phylogeny as inferred from heat shock protein 90 and ribosomal gene sequences.** *PLoS One* 2010, **5**(10):e13220.
262. Kitamura-Abe S, Itoh H, Washio T, Tsutsumi A, Tomita M: **Characterization of the splice sites in GT-AG and GC-AG introns in higher eukaryotes using full-length cDNAs.** *J Bioinform Comput Biol* 2004, **2**(2):309-331.
263. Nilsen TW: **The spliceosome: no assembly required?** *Mol Cell* 2002, **9**(1):8-9.
264. Reddy R, Spector D, Henning D, Liu MH, Busch H: **Isolation and partial characterization of dinoflagellate U1-U6 small RNAs homologous to rat U small nuclear RNAs.** *J Biol Chem* 1983, **258**(22):13965-13969.
265. Alverca E, Franca S, Diaz de la Espina SM: **Topology of splicing and snRNP biogenesis in dinoflagellate nuclei.** *Biol Cell* 2006, **98**(12):709-720.
266. Zhang H, Hou Y, Miranda L, Campbell DA, Sturm NR, Gaasterland T, Lin S: **Spliced leader RNA trans-splicing in dinoflagellates.** *Proc Natl Acad Sci U S A* 2007, **104**(11):4618-4623.
267. Boothroyd JC, Cross GA: **Transcripts coding for variant surface glycoproteins of *Trypanosoma brucei* have a short, identical exon at their 5' end.** *Gene* 1982, **20**(2):281-289.
268. Agabian N: **Trans splicing of nuclear pre-mRNAs.** *Cell* 1990, **61**(7):1157-1160.
269. Douris V, Telford MJ, Averof M: **Evidence for multiple independent origins of trans-splicing in Metazoa.** *Mol Biol Evol*, **27**(3):684-693.
270. Hastings KE: **SL trans-splicing: easy come or easy go?** *Trends Genet* 2005, **21**(4):240-247.

271. Lasda EL, Blumenthal T: **Trans-splicing**. *Wiley Interdiscip Rev RNA*, **2**(3):417-434.
272. Vandenberghe AE, Meedel TH, Hastings KE: **mRNA 5'-leader trans-splicing in the chordates**. *Genes Dev* 2001, **15**(3):294-303.
273. Davis RE: **Surprising diversity and distribution of spliced leader RNAs in flatworms**. *Mol Biochem Parasitol* 1997, **87**(1):29-48.
274. Lin S, Zhang H, Zhuang Y, Tran B, Gill J: **Spliced leader-based metatranscriptomic analyses lead to recognition of hidden genomic features in dinoflagellates**. *Proc Natl Acad Sci U S A* 2010, **107**(46):20033-20038.
275. Jackson CJ, Waller RF: **A widespread and unusual RNA trans-splicing type in dinoflagellate mitochondria**. *PLoS One* 2013, **8**(2):e56777.
276. Hearne JL, Pitula JS: **Identification of two spliced leader RNA transcripts from *Perkinsus marinus***. *J Eukaryot Microbiol*, **58**(3):266-268.
277. Slamovits CH, Keeling PJ: **Widespread recycling of processed cDNAs in dinoflagellates**. *Curr Biol* 2008, **18**(13):R550-552.
278. Zhang H, Dungan CF, Lin S: **Introns, alternative splicing, spliced leader trans-splicing and differential expression of pcna and cyclin in *Perkinsus marinus***. *Protist* 2011, **162**(1):154-167.
279. Suntharalingam M, Wentz SR: **Peering through the pore: nuclear pore complex structure, assembly, and function**. *Dev Cell* 2003, **4**(6):775-789.
280. Vasu SK, Forbes DJ: **Nuclear pores and nuclear assembly**. *Curr Opin Cell Biol* 2001, **13**(3):363-375.
281. Fried H, Kutay U: **Nucleocytoplasmic transport: taking an inventory**. *Cell Mol Life Sci* 2003, **60**(8):1659-1688.
282. Frankel MB, Knoll LJ: **The ins and outs of nuclear trafficking: unusual aspects in apicomplexan parasites**. *DNA Cell Biol* 2009, **28**(6):277-284.
283. Kohler A, Hurt E: **Exporting RNA from the nucleus to the cytoplasm**. *Nat Rev Mol Cell Biol* 2007, **8**(10):761-773.
284. Zhang J, Sun X, Qian Y, LaDuca JP, Maquat LE: **At least one intron is required for the nonsense-mediated decay of triosephosphate isomerase mRNA: a possible link between nuclear splicing and cytoplasmic translation**. *Mol Cell Biol* 1998, **18**(9):5272-5283.
285. Zhang J, Sun X, Qian Y, Maquat LE: **Intron function in the nonsense-mediated decay of beta-globin mRNA: indications that pre-mRNA splicing in the nucleus can influence mRNA translation in the cytoplasm**. *RNA* 1998, **4**(7):801-815.
286. Isken O, Maquat LE: **Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function**. *Genes Dev* 2007, **21**(15):1833-1856.
287. Ito-Harashima S, Kuroha K, Tatematsu T, Inada T: **Translation of the poly(A) tail plays crucial roles in nonstop mRNA surveillance via translation repression and protein destabilization by proteasome in yeast**. *Genes Dev* 2007, **21**(5):519-524.
288. Wu S, Wang W, Kong X, Congdon LM, Yokomori K, Kirschner MW, Rice JC: **Dynamic regulation of the PR-Set7 histone methyltransferase is required for normal cell cycle progression**. *Genes Dev* 2010, **24**(22):2531-2542.
289. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: molecular evolutionary genetics analysis using maximum likelihood,**

- evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 2011, **28**(10):2731-2739.
290. Mathews M. B. S, N. and Hershey, J. W. B.: **Translation control in Biology and Medicine** In: *Origin and Principles of translational control*. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY; 2000: 1 - 4.
 291. Warner JR, Knopf PM, Rich A: **A multiple ribosomal structure in protein synthesis**. *Proc Natl Acad Sci U S A* 1963, **49**:122-129.
 292. Raue HA, Klootwijk J, Musters W: **Evolutionary conservation of structure and function of high molecular weight ribosomal RNA**. *Prog Biophys Mol Biol* 1988, **51**(2):77-129.
 293. Alksne LE, Anthony RA, Liebman SW, Warner JR: **An accuracy center in the ribosome conserved over 2 billion years**. *Proc Natl Acad Sci U S A* 1993, **90**(20):9538-9541.
 294. Manuell AL, Yamaguchi K, Haynes PA, Milligan RA, Mayfield SP: **Composition and structure of the 80S ribosome from the green alga *Chlamydomonas reinhardtii*: 80S ribosomes are conserved in plants and animals**. *J Mol Biol* 2005, **351**(2):266-279.
 295. Wool IG, Chan YL, Gluck A: **Structure and evolution of mammalian ribosomal proteins**. *Biochem Cell Biol* 1995, **73**(11-12):933-947.
 296. Ross CL, Patel RR, Mendelson TC, Ware VC: **Functional conservation between structurally diverse ribosomal proteins from *Drosophila melanogaster* and *Saccharomyces cerevisiae*: fly L23a can substitute for yeast L25 in ribosome assembly and function**. *Nucleic Acids Res* 2007, **35**(13):4503-4514.
 297. Rae PM: **The nature and processing of ribosomal ribonucleic acid in a dinoflagellate**. *J Cell Biol* 1970, **46**(1):106-113.
 298. Maroteaux L, Herzog M, Soyer-Gobillard MO: **Molecular organization of dinoflagellate ribosomal DNA: evolutionary implications of the deduced 5.8 S rRNA secondary structure**. *Biosystems* 1985, **18**(3-4):307-319.
 299. Small HJ, Shields JD, Reece KS, Bateman K, Stentiford GD: **Morphological and molecular characterization of *Hematodinium perezii* (Dinophyceae: Syndiniales), a dinoflagellate parasite of the harbour crab, *Liocarcinus depurator***. *J Eukaryot Microbiol* 2012, **59**(1):54-66.
 300. Barna M: **Ribosomes take control**. *Proc Natl Acad Sci U S A*, **110**(1):9-10.
 301. Keiper BD, Lamphear BJ, Deshpande AM, Jankowska-Anyszka M, Aamodt EJ, Blumenthal T, Rhoads RE: **Functional characterization of five eIF4E isoforms in *Caenorhabditis elegans***. *J Biol Chem* 2000, **275**(14):10590-10596.
 302. Kruse S, Zhong S, Bodi Z, Button J, Alcocer MJ, Hayes CJ, Fray R: **A novel synthesis and detection method for cap-associated adenosine modifications in mouse mRNA**. *Sci Rep*, **1**:126.
 303. Jagus R, Bachvaroff TR, Joshi B, Place AR: **Diversity of Eukaryotic Translational Initiation Factor eIF4E in Protists**. *Comp Funct Genomics* 2012, **2012**:134839.
 304. Kozak M: **Regulation of translation via mRNA structure in prokaryotes and eukaryotes**. *Gene* 2005, **361**:13-37.
 305. Cigan AM, Donahue TF: **Sequence and structural features associated with translational initiator regions in yeast--a review**. *Gene* 1987, **59**(1):1-18.

306. Hamilton R, Watanabe CK, de Boer HA: **Compilation and comparison of the sequence context around the AUG startcodons in *Saccharomyces cerevisiae* mRNAs.** *Nucleic Acids Res* 1987, **15**(8):3581-3593.
307. Joshi CP, Zhou H, Huang X, Chiang VL: **Context sequences of translation initiation codon in plants.** *Plant Mol Biol* 1997, **35**(6):993-1001.
308. Xia X, Holcik M: **Strong eukaryotic IRESs have weak secondary structure.** *PLoS One* 2009, **4**(1):e4136.
309. Van Eden ME, Byrd MP, Sherrill KW, Lloyd RE: **Demonstrating internal ribosome entry sites in eukaryotic mRNAs using stringent RNA test procedures.** *RNA* 2004, **10**(4):720-730.
310. Kozak M: **A second look at cellular mRNA sequences said to function as internal ribosome entry sites.** *Nucleic Acids Res* 2005, **33**(20):6593-6602.
311. Kozak M: **Lessons (not) learned from mistakes about translation.** *Gene* 2007, **403**(1-2):194-203.
312. Kozak M: **Some thoughts about translational regulation: forward and backward glances.** *J Cell Biochem* 2007, **102**(2):280-290.
313. Kahvejian A, Svitkin YV, Sukarieh R, M'Boutchou MN, Sonenberg N: **Mammalian poly(A)-binding protein is a eukaryotic translation initiation factor, which acts via multiple mechanisms.** *Genes Dev* 2005, **19**(1):104-113.
314. Hinnebusch AG: **Mechanism and Regulation of Initiator Methionyl-tRNA binding to Ribosomes:** CSHL Press; 2000.
315. Price NT, Mellor H, Craddock BL, Flowers KM, Kimball SR, Wilmer T, Jefferson LS, Proud CG: **eIF2B, the guanine nucleotide-exchange factor for eukaryotic initiation factor 2. Sequence conservation between the alpha, beta and delta subunits of eIF2B from mammals and yeast.** *Biochem J* 1996, **318** (Pt 2):637-643.
316. Rozen F, Edery I, Meerovitch K, Dever TE, Merrick WC, Sonenberg N: **Bidirectional RNA helicase activity of eucaryotic translation initiation factors 4A and 4F.** *Mol Cell Biol* 1990, **10**(3):1134-1144.
317. Hereford LM, Osley MA, Ludwig TR, 2nd, McLaughlin CS: **Cell-cycle regulation of yeast histone mRNA.** *Cell* 1981, **24**(2):367-375.
318. Mirande M: **Processivity of translation in the eukaryote cell: role of aminoacyl-tRNA synthetases.** *FEBS Lett*, **584**(2):443-447.
319. Woese CR, Olsen GJ, Ibba M, Soll D: **Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process.** *Microbiol Mol Biol Rev* 2000, **64**(1):202-236.
320. Lipman RS, Hou YM: **Aminoacylation of tRNA in the evolution of an aminoacyl-tRNA synthetase.** *Proc Natl Acad Sci U S A* 1998, **95**(23):13495-13500.
321. Richter JD, Sonenberg N: **Regulation of cap-dependent translation by eIF4E inhibitory proteins.** *Nature* 2005, **433**(7025):477-480.
322. Gebauer F, Hentze MW: **Molecular mechanisms of translational control.** *Nat Rev Mol Cell Biol* 2004, **5**(10):827-835.
323. Karim MM, Svitkin YV, Kahvejian A, De Crescenzo G, Costa-Mattioli M, Sonenberg N: **A mechanism of translational repression by competition of Paip2 with eIF4G for poly(A) binding protein (PABP) binding.** *Proc Natl Acad Sci U S A* 2006, **103**(25):9494-9499.

324. Hershey JW: **Protein phosphorylation controls translation rates.** *J Biol Chem* 1989, **264**(35):20823-20826.
325. Shaikhin SM, Smailov SK, Lee AV, Kozhanov EV, Iskakov BK: **Interaction of wheat germ translation initiation factor 2 with GDP and GTP.** *Biochimie* 1992, **74**(5):447-454.
326. Lageix S, Lanet E, Pouch-Pelissier MN, Espagnol MC, Robaglia C, Deragon JM, Pelissier T: **Arabidopsis eIF2alpha kinase GCN2 is essential for growth in stress conditions and is activated by wounding.** *BMC Plant Biol* 2008, **8**:134.
327. Gingras AC, Svitkin Y, Belsham GJ, Pause A, Sonenberg N: **Activation of the translational suppressor 4E-BP1 following infection with encephalomyocarditis virus and poliovirus.** *Proc Natl Acad Sci U S A* 1996, **93**(11):5578-5583.
328. Gingras AC, Raught B, Gygi SP, Niedzwiecka A, Miron M, Burley SK, Polakiewicz RD, Wyslouch-Cieszyńska A, Aebersold R, Sonenberg N: **Hierarchical phosphorylation of the translation inhibitor 4E-BP1.** *Genes Dev* 2001, **15**(21):2852-2864.
329. Gray NK, Hentze MW: **Iron regulatory protein prevents binding of the 43S translation pre-initiation complex to ferritin and eALAS mRNAs.** *EMBO J* 1994, **13**(16):3882-3891.
330. Hinnebusch AG: **Translational regulation of GCN4 and the general amino acid control of yeast.** *Annu Rev Microbiol* 2005, **59**:407-450.
331. Mittag M, Eckerskorn C, Strupat K, Hastings JW: **Differential translational initiation of lbp mRNA is caused by a 5' upstream open reading frame.** *FEBS Lett* 1997, **411**(2-3):245-250.
332. Holcik M, Sonenberg N: **Translational control in stress and apoptosis.** *Nat Rev Mol Cell Biol* 2005, **6**(4):318-327.
333. Roenneberg T, Hastings JW: **Are the effects of light on phase and period of the Gonyaulax clock mediated by different pathways?** *Photochemistry and photobiology* 1991, **53**(4):525-533.
334. Hastings JW, Sweeney BM: **The luminescent reaction in extracts of the marine dinoflagellate, Gonyaulax polyedra.** *Journal of cellular physiology* 1957, **49**(2):209-225.
335. Krieger N, Hastings JW: **Bioluminescence: pH activity profiles of related luciferase fractions.** *Science (New York, NY)* 1968, **161**(841):586-589.
336. Bode VC, Desa R, Hastings JW: **Daily Rhythm of Luciferin Activity in Gonyaulax polyedra.** *Science* 1963, **141**(3584):913-915.
337. Johnson CH, Roeber JF, Hastings JW: **Circadian changes in enzyme concentration account for rhythm of enzyme activity in gonyaulax.** *Science* 1984, **223**(4643):1428-1430.
338. Mittag M, Lee DH, Hastings JW: **Circadian expression of the luciferin-binding protein correlates with the binding of a protein to the 3' untranslated region of its mRNA.** *Proceedings of the National Academy of Sciences of the United States of America* 1994, **91**(12):5257-5261.
339. Lapointe M, Morse D: **Reassessing the role of a 3'-UTR-binding translational inhibitor in regulation of circadian bioluminescence rhythm in the dinoflagellate Gonyaulax.** *Biol Chem* 2008, **389**(1):13-19.

340. Mittag M, Kiaulehn S, Johnson CH: **The circadian clock in *Chlamydomonas reinhardtii*. What is it for? What is it similar to?** *Plant Physiol* 2005, **137**(2):399-409.
341. Milos P, Morse D, Hastings JW: **Circadian control over synthesis of many *Gonyaulax* proteins is at a translational level.** *Naturwissenschaften* 1990, **77**(2):87-89.
342. Fagan T, Morse D, Hastings J: **Circadian synthesis of a nuclear encoded chloroplast Glyceraldehyde-3-phosphate dehydrogenase in the dinoflagellate *Gonyaulax polyedra* is translationally controlled.** *Biochemistry* 1999, **38**:7689-7695.
343. Le QH, Jovine, R., Markovic, P. and Mors, D.: **Peridinin-Chlorophyll a-Protein Is not Implicated in the Photosynthesis Rhythm of the dinoflagellate *Gonyaulax* despite Circadian Regulation of its Translation.** *Biological Rhythm Research* 2001, **32**(5):579 - 594.
344. Akimoto H, Kinumi T, Ohmiya Y: **Circadian rhythm of a TCA cycle enzyme is apparently regulated at the translational level in the dinoflagellate *Lingulodinium polyedrum*.** *J Biol Rhythms* 2005, **20**(6):479-489.
345. Wang DZ, Gao Y, Lin L, Hong HS: **Comparative proteomic analysis reveals proteins putatively involved in toxin biosynthesis in the marine dinoflagellate *Alexandrium catenella*.** *Mar Drugs* 2013, **11**(1):213-232.
346. Wang DZ, Li C, Zhang Y, Wang YY, He ZP, Lin L, Hong HS: **Quantitative proteomic analysis of differentially expressed proteins in the toxicity-lost mutant of *Alexandrium catenella* (Dinophyceae) in the exponential phase.** *J Proteomics* 2012, **75**(18):5564-5577.
347. Markovic P, Roenneberg T, Morse D: **Phased protein synthesis at several circadian times does not change protein levels in *Gonyaulax*.** *J Biol Rhythms* 1996, **11**(1):57-67.
348. Brunelle SA, Van Dolah FM: **Post-transcriptional regulation of S-phase genes in the dinoflagellate, *Karenia brevis*.** *J Eukaryot Microbiol*, **58**(4):373-382.
349. Zhang H, Hou Y, Lin S: **Isolation and characterization of proliferating cell nuclear antigen from the dinoflagellate *Pfiesteria piscicida*.** *J Eukaryot Microbiol* 2006, **53**(2):142-150.
350. Blais A, Dynlacht BD: **Hitting their targets: an emerging picture of E2F and cell cycle control.** *Curr Opin Genet Dev* 2004, **14**(5):527-532.
351. Mihailovich M, Militti C, Gabaldon T, Gebauer F: **Eukaryotic cold shock domain proteins: highly versatile regulators of gene expression.** *Bioessays* 2010, **32**(2):109-118.
352. Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Segurens B, Daubin V, Anthouard V, Aïach N *et al*: **Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*.** *Nature* 2006, **444**(7116):171-178.
353. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC: **Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*.** *Nature* 1998, **391**(6669):806-811.
354. Lee RC, Ambros V: **An extensive class of small RNAs in *Caenorhabditis elegans*.** *Science* 2001, **294**(5543):862-864.

355. Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B, Bartel DP: **MicroRNAs in plants.** *Genes Dev* 2002, **16**(13):1616-1626.
356. Liu J, Carmell MA, Rivas FV, Marsden CG, Thomson JM, Song JJ, Hammond SM, Joshua-Tor L, Hannon GJ: **Argonaute2 is the catalytic engine of mammalian RNAi.** *Science* 2004, **305**(5689):1437-1441.
357. Vasudevan S, Tong Y, Steitz JA: **Switching from repression to activation: microRNAs can up-regulate translation.** *Science* 2007, **318**(5858):1931-1934.
358. Sijen T, Fleenor J, Simmer F, Thijssen KL, Parrish S, Timmons L, Plasterk RH, Fire A: **On the role of RNA amplification in dsRNA-triggered gene silencing.** *Cell* 2001, **107**(4):465-476.
359. Cerutti H, Casas-Mollano JA: **On the origin and functions of RNA-mediated silencing: from protists to man.** *Curr Genet* 2006, **50**(2):81-99.
360. Baum J, Papenfuss AT, Mair GR, Janse CJ, Vlachou D, Waters AP, Cowman AF, Crabb BS, de Koning-Ward TF: **Molecular genetics and comparative genomics reveal RNAi is not functional in malaria parasites.** *Nucleic Acids Res* 2009, **37**(11):3788-3798.
361. Gunasekera AM, Patankar S, Schug J, Eisen G, Kissinger J, Roos D, Wirth DF: **Widespread distribution of antisense transcripts in the Plasmodium falciparum genome.** *Mol Biochem Parasitol* 2004, **136**(1):35-42.
362. De Riso V, Raniello R, Maumus F, Rogato A, Bowler C, Falciatore A: **Gene silencing in the marine diatom Phaeodactylum tricornutum.** *Nucleic Acids Res* 2009, **37**(14):e96.
363. Bazzini AA, Lee MT, Giraldez AJ: **Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish.** *Science*, **336**(6078):233-237.
364. Djuranovic S, Nahvi A, Green R: **miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay.** *Science*, **336**(6078):237-240.
365. Karve TM, Cheema AK: **Small changes huge impact: the role of protein posttranslational modifications in cellular homeostasis and disease.** *J Amino Acids* 2011, **2011**:207691.
366. Khoury GA, Baliban RC, Floudas CA: **Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database.** *Sci Rep*, **1**.
367. Cohen P: **The regulation of protein function by multisite phosphorylation--a 25 year update.** *Trends Biochem Sci* 2000, **25**(12):596-601.
368. Holt LJ, Tuch BB, Villen J, Johnson AD, Gygi SP, Morgan DO: **Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution.** *Science* 2009, **325**(5948):1682-1686.
369. Kersten B, Agrawal GK, Durek P, Neigenfind J, Schulze W, Walther D, Rakwal R: **Plant phosphoproteomics: an update.** *Proteomics* 2009, **9**(4):964-988.
370. Wang DZ, Li C, Xie ZX, Dong HP, Lin L, Hong HS: **Homology-Driven Proteomics of Dinoflagellates with Unsequenced Genomes Using MALDI-TOF/TOF and Automated De Novo Sequencing.** *Evid Based Complement Alternat Med* 2011, **2011**:471020.

371. Kusakina J, Dodd AN: **Phosphorylation in the plant circadian system.** *Trends Plant Sci* 2012, **17**(10):575-583.
372. Schafmeier T, Kaldi K, Diernfellner A, Mohr C, Brunner M: **Phosphorylation-dependent maturation of Neurospora circadian clock protein from a nuclear repressor toward a cytoplasmic activator.** *Genes Dev* 2006, **20**(3):297-306.
373. Dodd AN, Gardner MJ, Hotta CT, Hubbard KE, Dalchau N, Love J, Assie JM, Robertson FC, Jakobsen MK, Goncalves J *et al*: **The Arabidopsis circadian clock incorporates a cADPR-based feedback loop.** *Science* 2007, **318**(5857):1789-1792.
374. Edwards KD, Anderson PE, Hall A, Salathia NS, Locke JC, Lynn JR, Straume M, Smith JQ, Millar AJ: **FLOWERING LOCUS C mediates natural variation in the high-temperature response of the Arabidopsis circadian clock.** *Plant Cell* 2006, **18**(3):639-650.
375. Prosperi E, Scovassi AI, Stivala LA, Bianchi L: **Proliferating cell nuclear antigen bound to DNA synthesis sites: phosphorylation and association with cyclin D1 and cyclin A.** *Exp Cell Res* 1994, **215**(2):257-262.
376. Liu B, Lo SC, Matton DP, Lang BF, Morse D: **Daily changes in the phosphoproteome of the dinoflagellate lingulodinium.** *Protist* 2012, **163**(5):746-754.
377. Hochstrasser M: **Origin and function of ubiquitin-like proteins.** *Nature* 2009, **458**(7237):422-429.
378. Glickman MH, Ciechanover A: **The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction.** *Physiol Rev* 2002, **82**(2):373-428.
379. Komander D, Rape M: **The ubiquitin code.** *Annu Rev Biochem*, **81**:203-229.
380. Rill RL, Livolant F, Aldrich HC, Davidson MW: **Electron microscopy of liquid crystalline DNA: direct evidence for cholesteric-like organization of DNA in dinoflagellate chromosomes.** *Chromosoma* 1989, **98**(4):280-286.
381. Frenkiel-Krispin D, Levin-Zaidman S, Shimoni E, Wolf SG, Wachtel EJ, Arad T, Finkel SE, Kolter R, Minsky A: **Regulated phase transitions of bacterial chromatin: a non-enzymatic pathway for generic DNA protection.** *EMBO J* 2001, **20**(5):1184-1191.
382. Levi-Setti R, Gavrilov KL, Rizzo PJ: **Divalent cation distribution in dinoflagellate chromosomes imaged by high-resolution ion probe mass spectrometry.** *Eur J Cell Biol* 2008, **87**(12):963-976.
383. Herzog M, Soyer MO: **Distinctive features of dinoflagellate chromatin. Absence of nucleosomes in a primitive species Prorocentrum micans E.** *Eur J Cell Biol* 1981, **23**(2):295-302.
384. Lajeunesse TC: **"Species" radiations of symbiotic dinoflagellates in the Atlantic and Indo-Pacific since the Miocene-Pliocene transition.** *Mol Biol Evol* 2005, **22**(3):570-581.
385. Holm-Hansen O: **Algae: amounts of DNA and organic carbon in single cells.** *Science* 1969, **163**(862):87-88.
386. Saldarriaga JF, McEan ML, Fast NM, Taylor FJR, Keeling PJ: **Multiple protein phylogenies show that Oxyrrhis marina and Perkinsus marinus are early branches of the dinoflagellate lineage.** *Int J Syst Evol Microbiol* 2003, **53**:355-365.

387. Herzog M, Soyer M-O: **Distinctive features of dinoflagellate chromatin. Absence of nucleosomes in a primitive species *Prorocentrum micans*.** *Eur J Cell Biol* 1981, **23**:295-302.
388. Kapros T, Robertson AJ, Waterborg JH: **Histone H3 transcript stability in alfalfa.** *Plant Mol Biol* 1995, **28**(5):901-914.
389. Tanimoto EY, Rost TL, Comai L: **DNA Replication-Dependent Histone H2A mRNA Expression in Pea Root Tips.** *Plant Physiol* 1993, **103**(4):1291-1297.
390. Taoka K, Ohtsubo N, Fujimoto Y, Mikami K, Meshi T, Iwabuchi M: **The modular structure and function of the wheat H1 promoter with S phase-specific activity.** *Plant Cell Physiol* 1998, **39**(3):294-306.
391. Mosammaparast N, Guo Y, Shabanowitz J, Hunt DF, Pemberton LF: **Pathways mediating the nuclear import of histones H3 and H4 in yeast.** *J Biol Chem* 2002, **277**(1):862-868.
392. Mosammaparast N, Jackson KR, Guo Y, Brame CJ, Shabanowitz J, Hunt DF, Pemberton LF: **Nuclear import of histone H2A and H2B is mediated by a network of karyopherins.** *J Cell Biol* 2001, **153**(2):251-262.
393. Fries T, Betz C, Sohn K, Caesar S, Schlenstedt G, Bailer SM: **A novel conserved nuclear localization signal is recognized by a group of yeast importins.** *J Biol Chem* 2007, **282**(27):19292-19301.
394. Krude T: **Chromatin assembly during DNA replication in somatic cells.** *Eur J Biochem* 1999, **263**(1):1-5.
395. Park YJ, Luger K: **The structure of nucleosome assembly protein 1.** *Proc Natl Acad Sci U S A* 2006, **103**(5):1248-1253.
396. Vaquero A, Loyola A, Reinberg D: **The constantly changing face of chromatin.** *Sci Aging Knowledge Environ* 2003, **2003**(14):RE4.
397. Bannister AJ, Kouzarides T: **Regulation of chromatin by histone modifications.** *Cell Res* 2011, **21**(3):381-395.
398. Hackett JD, Scheetz TE, Yoon HS, Soares MB, Bonaldo MF, Casavant TL, Bhattacharya D: **Insights into a dinoflagellate genome through expressed sequence tag analysis.** *BMC Genomics* 2005, **6**:80.
399. Guillard RR, Ryther JH: **Studies of marine planktonic diatoms. I. *Cyclotella nana* Hustedt, and *Detonula confervacea* (Cleve) Grun.** *Can J Microbiol* 1962, **8**:229-239.
400. Chavez-Blanco A, Segura-Pacheco B, Perez-Cardenas E, Taja-Chayeb L, Cetina L, Candelaria M, Cantu D, Gonzalez-Fierro A, Garcia-Lopez P, Zambrano P *et al*: **Histone acetylation and histone deacetylase activity of magnesium valproate in tumor and peripheral blood of patients with cervical cancer. A phase I study.** *Mol Cancer* 2005, **4**(1):22.
401. Shechter D, Dormann HL, Allis CD, Hake SB: **Extraction, purification and analysis of histones.** *Nat Protoc* 2007, **2**(6):1445-1457.
402. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M *et al*: **Phylogeny.fr: robust phylogenetic analysis for the non-specialist.** *Nucleic Acids Res* 2008, **36**(Web Server issue):W465-469.

403. Moreno Diaz de la Espina S, Alverca E, Cuadrado A, Franca S: **Organization of the genome and gene expression in a nuclear environment lacking histones and nucleosomes: the amazing dinoflagellates.** *Eur J Cell Biol* 2005, **84**(2-3):137-149.
404. Costas E, Goyanes V: **Architecture and evolution of dinoflagellate chromosomes: an enigmatic origin.** *Cytogenet Genome Res* 2005, **109**(1-3):268-275.
405. Shupe K, Rizzo PJ: **Nuclease Digestion of Chromatin from the Eukaryotic Algae *Olithodiscus luteus*, *Peridinium balticum*, and *Cryptothecodinium cohnii*.** *Journal of Eukaryotic Microbiology* 1983, **30**(3):599-606.
406. Rizzo PJ, Cox ER: **Histone occurrence in chromatin from *Peridinium balticum*, a binucleate dinoflagellate.** *Science* 1977, **198**(4323):1258-1260.
407. Tomas RN, Cox ER, Steidinger KA: ***Peridinium balticum* (Levander) Lemmermann, an unusual dinoflagellate with a mesocaryotic and an eucaryotic nucleus.** *Journal of Phycology* 1973, **9**(1):91-98.
408. Postberg J, Forcob S, Chang WJ, Lipps HJ: **The evolutionary history of histone H3 suggests a deep eukaryotic root of chromatin modifying mechanisms.** *BMC Evol Biol* 2010, **10**:259.
409. West MH, Bonner WM: **Histone 2A, a heteromorphous family of eight protein species.** *Biochemistry* 1980, **19**(14):3238-3245.
410. Thatcher TH, Gorovsky MA: **Phylogenetic analysis of the core histones H2A, H2B, H3, and H4.** *Nucleic Acid Research* 1994, **22**(2):174-179.
411. Baxevas AD, Landsman D: **Histone Sequence Database: a compilation of highly-conserved nucleoprotein sequences.** *Nucleic Acids Res* 1996, **24**(1):245-247.
412. Rogakou EP, Pilch DR, Orr AH, Ivanova VS, Bonner WM: **DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139.** *J Biol Chem* 1998, **273**(10):5858-5868.
413. Dhillon N, Oki M, Szyjka SJ, Aparicio OM, Kamakaka RT: **H2A.Z functions to regulate progression through the cell cycle.** *Mol Cell Biol* 2006, **26**(2):489-501.
414. Robin P, Fritsch L, Philipot O, Svinarchuk F, Ait-Si-Ali S: **Post-translational modifications of histones H3 and H4 associated with the histone methyltransferases Suv39h1 and G9a.** *Genome Biol* 2007, **8**(12):R270.
415. Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE, Jr., Hieter P, Vogelstein B, Kinzler KW: **Characterization of the yeast transcriptome.** *Cell* 1997, **88**(2):243-251.
416. Suganuma T, Workman JL: **Signals and Combinatorial Functions of Histone Modifications.** *Annu Rev Biochem* 2010.
417. Verdin E, Dequiedt F, Kasler HG: **Class II histone deacetylases: versatile regulators.** *Trends Genet* 2003, **19**(5):286-293.
418. Margueron R, Trojer P, Reinberg D: **The key to development: interpreting the histone code?** *Curr Opin Genet Dev* 2005, **15**(2):163-176.
419. Glozak MA, Sengupta N, Zhang X, Seto E: **Acetylation and deacetylation of non-histone proteins.** *Gene* 2005, **363**:15-23.
420. Frye RA: **Phylogenetic classification of prokaryotic and eukaryotic Sir2-like proteins.** *Biochem Biophys Res Commun* 2000, **273**(2):793-798.

421. Starai VJ, Celic I, Cole RN, Boeke JD, Escalante-Semerena JC: **Sir2-dependent activation of acetyl-CoA synthetase by deacetylation of active lysine.** *Science* 2002, **298**(5602):2390-2392.
422. Martzen MR, McCraith SM, Spinelli SL, Torres FM, Fields S, Grayhack EJ, Phizicky EM: **A biochemical genomics approach for identifying genes by the activity of their products.** *Science* 1999, **286**(5442):1153-1155.
423. Trievel RC, Beach BM, Dirk LM, Houtz RL, Hurley JH: **Structure and catalytic mechanism of a SET domain protein methyltransferase.** *Cell* 2002, **111**(1):91-103.
424. Pennini ME, Perrinet S, Dautry-Varsat A, Subtil A: **Histone methylation by NUP, a novel nuclear effector of the intracellular pathogen Chlamydia trachomatis.** *PLoS Pathog* 2010, **6**(7):e1000995.
425. Canela N, Rodriguez-Vilarrupla A, Estanyol JM, Diaz C, Pujol MJ, Agell N, Bachs O: **The SET protein regulates G2/M transition by modulating cyclin B-cyclin-dependent kinase 1 activity.** *J Biol Chem* 2003, **278**(2):1158-1164.
426. Kellogg DR, Kikuchi A, Fujii-Nakata T, Turck CW, Murray AW: **Members of the NAP/SET family of proteins interact specifically with B-type cyclins.** *J Cell Biol* 1995, **130**(3):661-673.
427. Kellogg DR, Murray AW: **NAP1 acts with Clb1 to perform mitotic functions and to suppress polar bud growth in budding yeast.** *J Cell Biol* 1995, **130**(3):675-685.
428. Wu RS, Bonner WM: **Separation of basal histone synthesis from S-phase histone synthesis in dividing cells.** *Cell* 1981, **27**(2 Pt 1):321-330.
429. Robbins E, Borun TW: **The cytoplasmic synthesis of histones in hela cells and its temporal relationship to DNA replication.** *Proc Natl Acad Sci U S A* 1967, **57**(2):409-416.
430. Ehinger A, Denison SH, May GS: **Sequence, organization and expression of the core histone genes of Aspergillus nidulans.** *Mol Gen Genet* 1990, **222**(2-3):416-424.
431. Heintz N, Sive HL, Roeder RG: **Regulation of human histone gene expression: kinetics of accumulation and changes in the rate of synthesis and in the half-lives of individual histone mRNAs during the HeLa cell cycle.** *Mol Cell Biol* 1983, **3**(4):539-550.
432. Gordon BR, Leggat W: **Symbiodinium-invertebrate symbioses and the role of metabolomics.** *Mar Drugs* 2010, **8**(10):2546-2568.
433. Glibert PM, Anderson, D.M., Gentien, P., Granéli, E. and Sellner, K.: **The global, complex phenomena of harmful algal blooms.** *Oceanography* 2005, **18**:132-141.
434. Wilson T, Hastings JW: **Bioluminescence.** *Annu Rev Cell Dev Biol* 1998, **14**:197-230.
435. Spector D (ed.): **Dinoflagellates.** New York: Academic Press; 1984.
436. Imboden MA, Laird PW, Affolter M, Seebeck T: **Transcription of the intergenic regions of the tubulin gene cluster of Trypanosoma brucei: evidence for a polycistronic transcription unit in a eukaryote.** *Nucleic Acids Res* 1987, **15**(18):7357-7368.

437. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet* 2009, **10**(1):57-63.
438. Wang Y, Jensen L, Hojrup P, Morse D: **Synthesis and degradation of dinoflagellate plastid-encoded psbA proteins are light-regulated, not circadian-regulated.** *Proc Natl Acad Sci USA* 2005, **102**(8):2844-2849.
439. Drummond A, Ashton B, Buxton S, Cheung M, Cooper A, Duran C, Field M, Heled J, Kearse M, Markowitz S *et al*: **Geneious** In., v5.4 edn; 2011.
440. Tanikawa N, Akimoto H, Ogoh K, Chun W, Ohmiya Y: **Expressed sequence tag analysis of the dinoflagellate *Lingulodinium polyedrum* during dark phase.** *Photochem Photobiol* 2004, **80**:31-35.
441. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**(1):25-29.
442. Okuda S, Yamada T, Hamajima M, Itoh M, Katayama T, Bork P, Goto S, Kanehisa M: **KEGG Atlas mapping for global analysis of metabolic pathways.** *Nucleic Acids Res* 2008, **36**(Web Server issue):W423-426.
443. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics* 2005, **21**(18):3674-3676.
444. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K *et al*: **The Pfam protein families database.** *Nucleic Acids Res* 2010, **38**(Database issue):D211-222.
445. Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M *et al*: **The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism.** *Science* 2004, **306**(5693):79-86.
446. Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Marechal-Drouard L *et al*: **The *Chlamydomonas* genome reveals the evolution of key animal and plant functions.** *Science* 2007, **318**(5848):245-250.
447. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**(17):3389-3402.
448. Nei M, Gojobori T: **Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions.** *Mol Biol Evol* 1986, **3**(5):418-426.
449. Feldmeyer B, Wheat CW, Krezdorn N, Rotter B, Pfenninger M: **Short read Illumina data for the de novo assembly of a non-model snail species transcriptome (*Radix balthica*, Basommatophora, Pulmonata), and a comparison of assembler performance.** *BMC Genomics* 2011, **12**:317.
450. Morse D, Salois P, Markovic P, Hastings JW: **A nuclear encoded form II rubisco in dinoflagellates.** *Science* 1995, **268**:1622-1624.
451. Fagan T, Hastings JW, Morse D: **The phylogeny of glyceraldehyde-3-phosphate dehydrogenase indicates lateral gene transfer from cryptomonads to dinoflagellates.** *J Mol Evol* 1998, **47**(6):633-639.

452. Hackett JD, Yoon HS, Soares MB, Bonaldo MF, Casavant TL, Scheetz TE, Nosenko T, Bhattacharya D: **Migration of the plastid genome to the nucleus in a peridinin dinoflagellate.** *Curr Biol* 2004, **14**(3):213-218.
453. Lukes J, Leander BS, Keeling PJ: **Cascades of convergent evolution: the corresponding evolutionary histories of euglenozoans and dinoflagellates.** *Proc Natl Acad Sci U S A* 2009, **106** Suppl 1:9963-9970.
454. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5**(7):621-628.
455. Parsons M, Nelson RG, Watkins KP, Agabian N: **Trypanosome mRNAs share a common 5' spliced leader sequence.** *Cell* 1984, **38**(1):309-316.
456. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bahler J: **Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution.** *Nature* 2008, **453**(7199):1239-1243.
457. Ameur A, Zaghlool A, Halvardson J, Wetterbom A, Gyllenstein U, Cavelier L, Feuk L: **Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain.** *Nat Struct Mol Biol* 2011, **18**(12):1435-1440.
458. Lowe CD, Mello LV, Samatar N, Martin LE, Montagnes DJ, Watts PC: **The transcriptome of the novel dinoflagellate *Oxyrrhis marina* (Alveolata: Dinophyceae): response to salinity examined by 454 sequencing.** *BMC Genomics* 2011, **12**:519.
459. Reichman JR, Wilcox TP, Vize PD: **PCP gene family in *Symbiodinium* from *Hippopus hippopus*: low levels of concerted evolution, isoform diversity, and spectral tuning of chromophores.** *Mol Biol Evol* 2003, **20**(12):2143-2154.
460. Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP *et al*: **The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes.** *Nature* 2008, **456**(7219):239-244.
461. Salois P, Morse D: **Characterization and molecular phylogeny of a protein kinase cDNA from the dinoflagellate *Gonyaulax*.** *J Phycology* 1997, **33**:1063-1072.
462. McEwan M, Humayun R, Slamovits CH, Keeling PJ: **Nuclear genome sequence survey of the Dinoflagellate *Heterocapsa triquetra*.** *J Eukaryot Microbiol* 2008, **55**(6):530-535.
463. Takahashi N, Kaji H, Yanagida M, Hayano T, Isobe T: **Proteomics: advanced technology for the analysis of cellular function.** *J Nutr* 2003, **133**(6 Suppl 1):2090S-2096S.
464. Witze ES, Old WM, Resing KA, Ahn NG: **Mapping protein post-translational modifications with mass spectrometry.** *Nat Methods* 2007, **4**(10):798-806.
465. Smith JC, Figeys D: **Recent developments in mass spectrometry-based quantitative phosphoproteomics.** *Biochem Cell Biol* 2008, **86**(2):137-148.
466. Kersten B, Agrawal GK, Iwahashi H, Rakwal R: **Plant phosphoproteomics: a long road ahead.** *Proteomics* 2006, **6**(20):5517-5528.
467. Schulze WX, Usadel B: **Quantitation in mass-spectrometry-based proteomics.** *Annu Rev Plant Biol* 2010, **61**:491-516.

468. Jolma IW, Laerum OD, Lillo C, Ruoff P: **Circadian oscillators in eukaryotes.** *Wiley Interdiscip Rev Syst Biol Med* 2010, **2**(5):533-549.
469. Iliev D, Voytsekh O, Schmidt EM, Fiedler M, Nykytenko A, Mittag M: **A heteromeric RNA-binding protein is involved in maintaining acrophase and period of the circadian clock.** *Plant Physiol* 2006, **142**(2):797-806.
470. Morf J, Rey G, Schneider K, Stratmann M, Fujita J, Naef F, Schibler U: **Cold-inducible RNA-binding protein modulates circadian gene expression posttranscriptionally.** *Science* 2012, **338**(6105):379-383.
471. Mizoguchi T, Putterill J, Ohkoshi Y: **Kinase and phosphatase: the cog and spring of the circadian clock.** *Int Rev Cytol* 2006, **250**:47-72.
472. Lu SX, Liu H, Knowles SM, Li J, Ma L, Tobin EM, Lin C: **A role for protein kinase casein kinase2 alpha-subunits in the Arabidopsis circadian clock.** *Plant Physiol* 2011, **157**(3):1537-1545.
473. Ouyang Y, Andersson CR, Kondo T, Golden SS, Johnson CH: **Resonating circadian clocks enhance fitness in cyanobacteria.** *Proc Natl Acad Sci U S A* 1998, **95**(15):8660-8664.
474. Coste H, Brevet A, Plateau P, Blanquet S: **Non-adenylylated bis(5'-nucleosidyl) tetraphosphates occur in *Saccharomyces cerevisiae* and in *Escherichia coli* and accumulate upon temperature shift or exposure to cadmium.** *J Biol Chem* 1987, **262**(25):12096-12103.
475. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L *et al*: **InterPro: the integrative protein signature database.** *Nucleic Acids Res* 2009, **37**(Database issue):D211-215.
476. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C *et al*: **Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data.** *Bioinformatics* 2012, **28**(12):1647-1649.
477. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S: **The protein kinase complement of the human genome.** *Science* 2002, **298**(5600):1912-1934.
478. Bradham CA, Foltz KR, Beane WS, Arnone MI, Rizzo F, Coffman JA, Mushegian A, Goel M, Morales J, Geneviere AM *et al*: **The sea urchin kinome: a first look.** *Dev Biol* 2006, **300**(1):180-193.
479. Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, Wortman JR, Badger JH, Ren Q, Amedeo P, Jones KM *et al*: **Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote.** *PLoS Biol* 2006, **4**(9):e286.
480. Xue Y, Ren J, Gao X, Jin C, Wen L, Yao X: **GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy.** *Mol Cell Proteomics* 2008, **7**(9):1598-1608.
481. Davis MA, Hinerfeld D, Joseph S, Hui YH, Huang NH, Leszyk J, Rutherford-Bethard J, Tam SW: **Proteomic analysis of rat liver phosphoproteins after treatment with protein kinase inhibitor H89 (N-(2-[p-bromocinnamylamino]-ethyl)-5-isoquinolinesulfonamide).** *J Pharmacol Exp Ther* 2006, **318**(2):589-595.
482. Cerny M, Dycka F, Bobal'ova J, Brzobohaty B: **Early cytokinin response proteins and phosphoproteins of *Arabidopsis thaliana* identified by proteome and phosphoproteome profiling.** *J Exp Bot* 2011, **62**(3):921-937.

483. Laugesen S, Messinese E, Hem S, Pichereaux C, Grat S, Ranjeva R, Rossignol M, Bono JJ: **Phosphoproteins analysis in plants: a proteomic approach.** *Phytochemistry* 2006, **67**(20):2208-2214.
484. Mann M, Ong SE, Gronborg M, Steen H, Jensen ON, Pandey A: **Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome.** *Trends Biotechnol* 2002, **20**(6):261-268.
485. Wagner V, Gessner G, Heiland I, Kaminski M, Hawat S, Scheffler K, Mittag M: **Analysis of the phosphoproteome of *Chlamydomonas reinhardtii* provides new insights into various cellular pathways.** *Eukaryot Cell* 2006, **5**(3):457-468.
486. Sugiyama N, Nakagami H, Mochida K, Daudi A, Tomita M, Shirasu K, Ishihama Y: **Large-scale phosphorylation mapping reveals the extent of tyrosine phosphorylation in *Arabidopsis*.** *Mol Syst Biol* 2008, **4**:193.
487. Sugiyama N, Masuda T, Shinoda K, Nakamura A, Tomita M, Ishihama Y: **Phosphopeptide enrichment by aliphatic hydroxy acid-modified metal oxide chromatography for nano-LC-MS/MS in proteomics applications.** *Mol Cell Proteomics* 2007, **6**(6):1103-1109.
488. Molina H, Horn DM, Tang N, Mathivanan S, Pandey A: **Global proteomic profiling of phosphopeptides using electron transfer dissociation tandem mass spectrometry.** *Proc Natl Acad Sci U S A* 2007, **104**(7):2199-2204.
489. Olsen JV, Blagoev B, Gnäd F, Macek B, Kumar C, Mortensen P, Mann M: **Global, in vivo, and site-specific phosphorylation dynamics in signaling networks.** *Cell* 2006, **127**(3):635-648.
490. Nakagami H, Sugiyama N, Mochida K, Daudi A, Yoshida Y, Toyoda T, Tomita M, Ishihama Y, Shirasu K: **Large-scale comparative phosphoproteomics identifies conserved phosphorylation sites in plants.** *Plant Physiol* 2010, **153**(3):1161-1174.
491. Muller WE: **Review: How was metazoan threshold crossed? The hypothetical Urmetazoa.** *Comp Biochem Physiol A Mol Integr Physiol* 2001, **129**(2-3):433-460.
492. Kloss B, Price JL, Saez L, Blau J, Rothenfluh A, Wesley CS, Young MW: **The *Drosophila* clock gene double-time encodes a protein closely related to human casein kinase Iepsilon.** *Cell* 1998, **94**(1):97-107.
493. Martinek S, Inonog S, Manoukian AS, Young MW: **A role for the segment polarity gene shaggy/GSK-3 in the *Drosophila* circadian clock.** *Cell* 2001, **105**(6):769-779.
494. Lamia KA, Sachdeva UM, DiTacchio L, Williams EC, Alvarez JG, Egan DF, Vasquez DS, Juguilon H, Panda S, Shaw RJ *et al*: **AMPK regulates the circadian clock by cryptochrome phosphorylation and degradation.** *Science* 2009, **326**(5951):437-440.
495. Meggio F, Pinna LA: **One-thousand-and-one substrates of protein kinase CK2?** *FASEB J* 2003, **17**(3):349-368.
496. Nagy E, Rigby WF: **Glyceraldehyde-3-phosphate dehydrogenase selectively binds AU-rich RNA in the NAD(+)-binding region (Rossmann fold).** *J Biol Chem* 1995, **270**(6):2755-2763.
497. Dollenmaier G, Weitz M: **Interaction of glyceraldehyde-3-phosphate dehydrogenase with secondary and tertiary RNA structural elements of the hepatitis A virus 3' translated and non-translated regions.** *J Gen Virol* 2003, **84**(Pt 2):403-414.

498. Kline-Jonakin KG, Barrett-Wilt GA, Sussman MR: **Quantitative plant phosphoproteomics.** *Curr Opin Plant Biol* 2011, **14**(5):507-511.
499. Tetlow IJ, Wait R, Lu Z, Akkasaeng R, Bowsher CG, Esposito S, Kosar-Hashemi B, Morell MK, Emes MJ: **Protein phosphorylation in amyloplasts regulates starch branching enzyme activity and protein-protein interactions.** *Plant Cell* 2004, **16**(3):694-708.
500. Schneegurt MA, Sherman DM, Nayar S, Sherman LA: **Oscillating behavior of carbohydrate granule formation and dinitrogen fixation in the cyanobacterium *Cyanothece* sp. strain ATCC 51142.** *J Bacteriol* 1994, **176**(6):1586-1597.
501. Adeli K: **Translational control mechanisms in metabolic regulation: critical role of RNA binding proteins, microRNAs, and cytoplasmic RNA granules.** *Am J Physiol Endocrinol Metab* 2011, **301**(6):E1051-1064.
502. Wilkie GS, Dickson KS, Gray NK: **Regulation of mRNA translation by 5'- and 3'-UTR-binding factors.** *Trends Biochem Sci* 2003, **28**(4):182-188.
503. Ptacek J, Snyder M: **Charging it up: global analysis of protein phosphorylation.** *Trends Genet* 2006, **22**(10):545-554.
504. Dreyfuss G, Swanson MS, Pinol-Roma S: **Heterogeneous nuclear ribonucleoprotein particles and the pathway of mRNA formation.** *Trends Biochem Sci* 1988, **13**(3):86-91.
505. Chambers JC, Kenan D, Martin BJ, Keene JD: **Genomic structure and amino acid sequence domains of the human La autoantigen.** *J Biol Chem* 1988, **263**(34):18043-18051.
506. Sachs AB, Davis RW, Kornberg RD: **A single domain of yeast poly(A)-binding protein is necessary and sufficient for RNA binding and cell viability.** *Mol Cell Biol* 1987, **7**(9):3268-3276.
507. Ogura K, Kishimoto N, Mitani S, Gengyo-Ando K, Kohara Y: **Translational control of maternal glp-1 mRNA by POS-1 and its interacting protein SPN-4 in *Caenorhabditis elegans*.** *Development* 2003, **130**(11):2495-2503.
508. Lee EK, Kim W, Tominaga K, Martindale JL, Yang X, Subaran SS, Carlson OD, Mercken EM, Kulkarni RN, Akamatsu W *et al*: **RNA-binding protein HuD controls insulin translation.** *Mol Cell* 2012, **45**(6):826-835.
509. Gebauer F, Corona DF, Preiss T, Becker PB, Hentze MW: **Translational control of dosage compensation in *Drosophila* by Sex-lethal: cooperative silencing via the 5' and 3' UTRs of msl-2 mRNA is independent of the poly(A) tail.** *EMBO J* 1999, **18**(21):6146-6154.
510. Reyes R, Izquierdo JM: **The RNA-binding protein PTB exerts translational control on 3'-untranslated region of the mRNA for the ATP synthase beta-subunit.** *Biochem Biophys Res Commun* 2007, **357**(4):1107-1112.
511. Garcia-Mayoral MF, Hollingworth D, Masino L, Diaz-Moreno I, Kelly G, Gherzi R, Chou CF, Chen CY, Ramos A: **The structure of the C-terminal KH domains of KSRP reveals a noncanonical motif important for mRNA degradation.** *Structure* 2007, **15**(4):485-498.
512. Paquin N, Menade M, Poirier G, Donato D, Drouet E, Chartrand P: **Local activation of yeast ASH1 mRNA translation through phosphorylation of Khd1p by the casein kinase Yck1p.** *Mol Cell* 2007, **26**(6):795-809.

513. Zhang B, Gallegos M, Puoti A, Durkin E, Fields S, Kimble J, Wickens MP: **A conserved RNA-binding protein that regulates sexual fates in the *C. elegans* hermaphrodite germ line.** *Nature* 1997, **390**(6659):477-484.
514. Wang X, McLachlan J, Zamore PD, Hall TM: **Modular recognition of RNA by a human pumilio-homology domain.** *Cell* 2002, **110**(4):501-512.
515. Wang X, Zamore PD, Hall TM: **Crystal structure of a Pumilio homology domain.** *Mol Cell* 2001, **7**(4):855-865.
516. Deng Y, Singer RH, Gu W: **Translation of ASH1 mRNA is repressed by Puf6p-Fun12p/eIF5B interaction and released by CK2 phosphorylation.** *Genes Dev* 2008, **22**(8):1037-1050.
517. Lurin C, Andres C, Aubourg S, Bellaoui M, Bitton F, Bruyere C, Caboche M, Debast C, Gualberto J, Hoffmann B *et al*: **Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis.** *Plant Cell* 2004, **16**(8):2089-2103.
518. Ermolenko DN, Makhataдзе GI: **Bacterial cold-shock proteins.** *Cell Mol Life Sci* 2002, **59**(11):1902-1913.
519. Sasaki K, Imai R: **Pleiotropic roles of cold shock domain proteins in plants.** *Front Plant Sci* 2011, **2**:116.
520. Tanaka T, Mega R, Kim K, Shinkai A, Masui R, Kuramitsu S, Nakagawa N: **A non-cold-inducible cold shock protein homolog mainly contributes to translational control under optimal growth conditions.** *FEBS J* 2012.
521. Murray MT, Schiller DL, Franke WW: **Sequence analysis of cytoplasmic mRNA-binding proteins of *Xenopus* oocytes identifies a family of RNA-binding proteins.** *Proc Natl Acad Sci U S A* 1992, **89**(1):11-15.
522. Ranjan M, Tafuri SR, Wolffe AP: **Masking mRNA from translation in somatic cells.** *Genes Dev* 1993, **7**(9):1725-1736.
523. Nekrasov MP, Ivshina MP, Chernov KG, Kovrigina EA, Evdokimova VM, Thomas AA, Hershey JW, Ovchinnikov LP: **The mRNA-binding protein YB-1 (p50) prevents association of the eukaryotic initiation factor eIF4G with mRNA and inhibits protein synthesis at the initiation stage.** *J Biol Chem* 2003, **278**(16):13936-13943.
524. Lam CM, Yeung PK, Wong JT: **Monitoring cytosolic calcium in the dinoflagellate *Cryptothecodinium cohnii* with calcium orange-AM.** *Plant Cell Physiol* 2005, **46**(6):1021-1027.
525. von Dassow P, Latz MI: **The role of Ca(2+) in stimulated bioluminescence of the dinoflagellate *Lingulodinium polyedrum*.** *J Exp Biol* 2002, **205**(Pt 19):2971-2986.
526. Filhol O, Cochet C: **Protein kinase CK2 in health and disease: Cellular functions of protein kinase CK2: a dynamic affair.** *Cell Mol Life Sci* 2009, **66**(11-12):1830-1839.
527. Haapala OK, Soyer MO: **Electron microscopy of whole-mounted chromosomes of the dinoflagellate *Gyrodinium cohnii*.** *Hereditas* 1974, **78**(1):146-150.
528. Rizzo PJ: **Histones in protistan evolution.** *Biosystems* 1985, **18**(3-4):249-262.
529. Anderson DM, Cembella AD, Hallegraeff GM: **Progress in understanding harmful algal blooms: paradigm shifts and new technologies for research, monitoring, and management.** *Ann Rev Mar Sci* 2012, **4**:143-176.

530. Wang DZ: **Neurotoxins from marine dinoflagellates: a brief review.** *Mar Drugs* 2008, **6**(2):349-371.
531. Keafer BA, Churchill, J.H., Anderson, D.M.: **Blooms of the toxic dinoflagellate, *Alexandrium fundyense*, in the Casco Bay region of the western Gulf of Maine: Advection from offshore source populations and interactions with the Kennebec River plume.** *Deep-Sea Res II* 2005, **52** (19-21):2631-2655.
532. Ribeiro S, Berge T, Lundholm N, Andersen TJ, Abrantes F, Ellegaard M: **Phytoplankton growth after a century of dormancy illuminates past resilience to catastrophic darkness.** *Nat Commun* 2011, **2**:311.
533. Lewis JaH, R.: **Lingulodinium polyedrum (*Gonyaulax polyedra*) a blooming dinoflagellate.**, vol. 35. London: UCL Press 1997.
534. Kokinos JP, Anderson, D. M.: **Morphological development of resting cysts in cultures of the marine dinoflagellate *Lingulodinium polyedrum* (= *L. machaerophorum*).** *Palynology* 1995, **19**(143-166).
535. Tslm ST, Wong JT, Wong YH: **CGP 52608-induced cyst formation in dinoflagellates: possible involvement of a nuclear receptor for melatonin.** *J Pineal Res* 1996, **21**(2):101-107.
536. Balzer I, Hardeland R: **Photoperiodism and effects of indoleamines in a unicellular alga, *Gonyaulax polyedra*.** *Science* 1991, **253**(5021):795-797.
537. Bravo I, Figueroab, R.I., Garcésb, E. Fraga, S., Massanet, A.: **The intricacies of dinoflagellate pellicle cysts: The example of *Alexandrium minutum* cysts from a bloom-recurrent area (Bay of Baiona, NW Spain).** *Deep Sea Research Part II: Topical Studies in Oceanography* 2010, **57**(3-4):166-174.
538. Behrmann G, Hardeland, R.: **Ultrastructural characterization of asexual cysts of *Gonyaulax polyedra* Stein (Dinoflagellata).** *Protoplasma* 1995, **185**(1-2):22-27.
539. Price AM, Pospelova, V.: **High-resolution sediment trap study of organic-walled dinoflagellate cyst production and biogenic silica flux in Saanich Inlet (BC, Canada).** *Marine Micropaleontology* 2011, **80**:18-43.
540. Horn G, Hofweber R, Kremer W, Kalbitzer HR: **Structure and function of bacterial cold shock proteins.** *Cell Mol Life Sci* 2007, **64**(12):1457-1470.
541. Thomashow MF: **Molecular basis of plant cold acclimation: insights gained from studying the CBF cold response pathway.** *Plant Physiol* 2010, **154**(2):571-577.
542. Hastings JW: **Cellular and Molecular Mechanisms of Circadian Regulation in the Unicellular Dinoflagellate *Gonyaulax polyedra*.** *Handbook of Behavioral Neurobiology* 2001, **12**:321-334.
543. Morse D, Pappenheimer AM, Jr., Hastings JW: **Role of a luciferin-binding protein in the circadian bioluminescent reaction of *Gonyaulax polyedra*.** *J Biol Chem* 1989, **264**(20):11822-11826.
544. Morse D, Mittag M: **Dinoflagellate luciferin-binding protein.** *Methods Enzymol* 2000, **305**:258-276.
545. Anderson DM, Keafer BA: **An endogenous annual clock in the toxic marine dinoflagellate *Gonyaulax tamarensis*.** *Nature* 1987, **325**(6105):616-617.
546. Njus D, McMurry, L., Hastings, J.W.: **Conditionality of Circadian Rhythmicity: Synergistic Action of Light and Temperature.** *J Comp Physiol* 1977, **117**:335-344.

547. Bieniawska Z, Espinoza C, Schlereth A, Sulpice R, Hinch DK, Hannah MA: **Disruption of the Arabidopsis circadian clock is responsible for extensive variation in the cold-responsive transcriptome.** *Plant Physiol* 2008, **147**(1):263-279.
548. Seo PJ, Park MJ, Lim MH, Kim SG, Lee M, Baldwin IT, Park CM: **A self-regulatory circuit of CIRCADIAN CLOCK-ASSOCIATED1 underlies the circadian clock regulation of temperature responses in Arabidopsis.** *Plant Cell*, **24**(6):2427-2442.
549. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al*: **Full-length transcriptome assembly from RNA-Seq data without a reference genome.** *Nat Biotechnol* 2011, **29**(7):644-652.
550. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al*: **Full-length transcriptome assembly from RNA-Seq data without a reference genome.** *Nat Biotechnol*, **29**(7):644-652.
551. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**(14):1754-1760.
552. Wang L, Feng Z, Wang X, Zhang X: **DEGseq: an R package for identifying differentially expressed genes from RNA-seq data.** *Bioinformatics* 2010, **26**(1):136-138.
553. Feinberg AP, Vogelstein B: **A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity.** *Anal Biochem* 1983, **132**(1):6-13.
554. Lee BH, Henderson DA, Zhu JK: **The Arabidopsis cold-responsive transcriptome and its regulation by ICE1.** *Plant Cell* 2005, **17**(11):3155-3175.
555. Anderson DM, Wall, D.: **Potential importance of benthic cysts of *Gonyaulax tamarensis* and *G. excavata* in initiating toxic dinoflagellate blooms.** *J Phycol* 1978, **14**:224-234.
556. Fistarol GO, Legrand C, Rengefors K, Graneli E: **Temporary cyst formation in phytoplankton: a response to allelopathic competitors?** *Environ Microbiol* 2004, **6**(8):791-798.
557. Rengefors K, Karlsson, I. Hansson, L. A.: **Algal cyst dormancy: a temporal escape from herbivory.** *Proc Biol Sci* 1998, **265**(1403):1353-1358.
558. Saito K, Drgon T, Krupatkina DN, Drgonova J, Terlizzi DE, Mercer N, Vasta GR: **Effect of biotic and abiotic factors on in vitro proliferation, encystment, and excystment of *Pfiesteria piscicida*.** *Appl Environ Microbiol* 2007, **73**(20):6410-6420.
559. Evitt WR: **Sporopollenin dinoflagellate cysts: their morphology and interpretation.** Dallas: American Association of Stratigraphic Palynologists.; 1985.
560. Anderson DM: **Effects of temperature conditioning on development and germination of (*Gonyaulax tomarensis* Dinophyceae) hypnozygotes.** *J phycol* 1980, **16**:166-172.
561. Al-Fageeh MB, Smales CM: **Control and regulation of the cellular responses to cold shock: the responses in yeast and mammalian systems.** *Biochem J* 2006, **397**(2):247-259.
562. Phadtare S: **Unwinding activity of cold shock proteins and RNA metabolism.** *RNA Biol* 2011, **8**(3):394-397.

563. Gualerzi CO, Giuliodori AM, Pon CL: **Transcriptional and post-transcriptional control of cold-shock genes.** *J Mol Biol* 2003, **331**(3):527-539.
564. Yan SP, Zhang QY, Tang ZC, Su WA, Sun WN: **Comparative proteomic analysis provides new insights into chilling stress responses in rice.** *Mol Cell Proteomics* 2006, **5**(3):484-496.
565. Rinalducci S, Egidi MG, Karimzadeh G, Jazii FR, Zolla L: **Proteomic analysis of a spring wheat cultivar in response to prolonged cold stress.** *Electrophoresis* 2011, **32**(14):1807-1818.
566. Rinalducci S, Egidi MG, Mahfoozi S, Godehkahriz SJ, Zolla L: **The influence of temperature on plant development in a vernalization-requiring winter wheat: A 2-DE based proteomic investigation.** *J Proteomics* 2011, **74**(5):643-659.
567. Neilson KA, Mariani M, Haynes PA: **Quantitative proteomic analysis of cold-responsive proteins in rice.** *Proteomics* 2011, **11**(9):1696-1706.
568. Xin H, Zhu W, Wang L, Xiang Y, Fang L, Li J, Sun X, Wang N, Londo JP, Li S: **Genome Wide Transcriptional Profile Analysis of Vitis amurensis and Vitis vinifera in Response to Cold Stress.** *PLoS One* 2013, **8**(3):e58740.
569. Hannah MA, Heyer AG, Hinch DK: **A global survey of gene regulation during cold acclimation in Arabidopsis thaliana.** *PLoS Genet* 2005, **1**(2):e26.
570. Chiba Y, Mineta K, Hirai MY, Suzuki Y, Kanaya S, Takahashi H, Onouchi H, Yamaguchi J, Naito S: **Changes in mRNA stability associated with cold stress in Arabidopsis cells.** *Plant Cell Physiol* 2012, **54**(2):180-194.
571. Yamanaka K, Inouye M: **Selective mRNA degradation by polynucleotide phosphorylase in cold shock adaptation in Escherichia coli.** *J Bacteriol* 2001, **183**(9):2808-2816.
572. Marzluff WF, Wagner EJ, Duronio RJ: **Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail.** *Nat Rev Genet* 2008, **9**(11):843-854.
573. Chen CY, Shyu AB: **Selective degradation of early-response-gene mRNAs: functional analyses of sequence features of the AU-rich elements.** *Mol Cell Biol* 1994, **14**(12):8471-8482.
574. Sarkar B, Xi Q, He C, Schneider RJ: **Selective degradation of AU-rich mRNAs promoted by the p37 AUF1 protein isoform.** *Mol Cell Biol* 2003, **23**(18):6685-6693.
575. Curatola AM, Nadal MS, Schneider RJ: **Rapid degradation of AU-rich element (ARE) mRNAs is activated by ribosome transit and blocked by secondary structure at any position 5' to the ARE.** *Mol Cell Biol* 1995, **15**(11):6331-6340.
576. Fan XC, Myer VE, Steitz JA: **AU-rich elements target small nuclear RNAs as well as mRNAs for rapid degradation.** *Genes Dev* 1997, **11**(19):2557-2568.
577. Jacobs J, Kuck U: **Function of chloroplast RNA-binding proteins.** *Cell Mol Life Sci* 2011, **68**(5):735-748.
578. Chen J, Tian L., Xu, H., Tian, D., Luo, Y., Ren, C., Yang, L., Shi, J.: **Cold-induced changes of protein and phosphoprotein expression patterns from rice roots as revealed by multiplex proteomic analysis.** *Plant Omics Journal* 2012, **5**(2):194-199.

579. Kawczynski W, Dhindsa, R.S.: **Alfalfa Nuclei Contain Cold-Responsive Phosphoproteins and Accumulate Heat-Stable Proteins during Cold Treatment of Seedlings.** *Plant Cell Physiol* 1996, **37**(8):1204-1210.
580. Dang CV, Traugh JA: **Phosphorylation of threonyl- and seryl-tRNA synthetase by cAMP-dependent protein kinase. A possible role in the regulation of P1, P4-bis(5'-adenosyl)-tetraphosphate (Ap4A) synthesis.** *J Biol Chem* 1989, **264**(10):5861-5865.
581. Harnett SP, Lowe G, Tansley G: **Mechanism of activation of phenylalanine and synthesis of P1, P4-bis(5'-adenosyl) tetraphosphate by yeast phenylalanyl-tRNA synthetase.** *Biochemistry* 1985, **24**(12):2908-2915.
582. Kobayashi Y, Kuratomi K: **The binding activities of proteins that bind Ap4A, an alarmone, are stimulated in the presence of ethanol or phosphatidylethanolamine.** *Biochem Biophys Res Commun* 1989, **160**(3):1379-1386.
583. Lee PC, Bochner BR, Ames BN: **Diadenosine 5',5'''-P1,P4-tetraphosphate and related adenylylated nucleotides in Salmonella typhimurium.** *J Biol Chem* 1983, **258**(11):6827-6834.
584. Guranowski A, Sillero MA, Sillero A: **Firefly luciferase synthesizes P1,P4-bis(5'-adenosyl)tetraphosphate (Ap4A) and other dinucleoside polyphosphates.** *FEBS Lett* 1990, **271**(1-2):215-218.
585. Knight H: **Calcium signaling during abiotic stress in plants.** *Int Rev Cytol* 2000, **195**:269-324.
586. Lee C, Etchegaray JP, Cagampang FR, Loudon AS, Reppert SM: **Posttranslational mechanisms regulate the mammalian circadian clock.** *Cell* 2001, **107**(7):855-867.
587. Hastings JW: **Chemistry, clones, and circadian control of the dinoflagellate bioluminescent system. The Marlene DeLuca memorial lecture.** *J Biolumin Chemilumin* 1989, **4**(1):12-19.
588. Wek RC, Jiang HY, Anthony TG: **Coping with stress: eIF2 kinases and translational control.** *Biochem Soc Trans* 2006, **34**(Pt 1):7-11.

